IL PROBLEMA DELL'INTERNO

Soggettività, prospettiva e il confine tra modello e mente

Prefazione

Questo contributo scaturisce da un'inquietudine che percorre obliquamente la filosofia, la matematica e le scienze cognitive: l'inquietudine di chi si avvede che i sistemi artificiali più raffinati mai costruiti dall'ingegno umano difettano di qualcosa che non si riesce a denominare con esattezza. Non è intelligenza ciò che manca — quella, in qualche misura tecnica, questi sistemi la possiedono, e spesso la esibiscono con una disinvoltura che lascia perplessi. Non è linguaggio — i modelli a larga scala parametrica conversano con una fluidità che sovente imbarazza l'interlocutore umano, costringendolo a interrogarsi se stia parlando con qualcosa o con qualcuno. Non è nemmeno conoscenza nel senso enciclopedico e cumulativo del vocabolo: nessuna biblioteca umana eguaglia l'estensione di ciò che un modello addestrato su miliardi di frammenti testuali ha sedimentato nei propri parametri.

Ciò che manca è qualcosa di più elementare e insieme più refrattario all'analisi: un *interno*. Un punto di vista. Una prospettiva che non discenda dalla rappresentazione del mondo, ma ne costituisca la condizione trascendentale di possibilità. Un luogo — non geometrico, non spaziale nel senso triviale del termine — dal quale il mondo *appare*, piuttosto che essere semplicemente descritto, catalogato o predetto.

La distinzione è sottile ma abissale. Una mappa rappresenta il territorio senza trovarsi da nessuna parte in particolare; eppure il viaggiatore che consulta quella mappa occupa sempre un punto su di essa, un *qui* che orienta ogni *là* e ogni *laggiù*. La mappa conosce il nord oggettivo; il viaggiatore conosce la propria destra e la propria sinistra. La mappa è terza persona assoluta: il viaggiatore è

prima persona irriducibile. Tra i due non intercorre solo una differenza di complessità o di risoluzione: vi è una differenza di categoria ontologica.

Il presente saggio si propone di interrogare questa differenza con gli strumenti che giudichiamo i più consoni all'altezza del problema: la teoria dei modelli e la matematica strutturale, la fenomenologia dell'incarnazione, la teoria dell'informazione e una riflessione serrata sull'architettura dei sistemi di intelligenza artificiale che oggi animano il dibattito scientifico e filosofico. L'obiettivo non è conseguire risposte definitive — sarebbe una pretesa sproporzionata rispetto alla natura di una questione che ha resistito a millenni di speculazione — bensì avanzare verso una formulazione più nitida e matematicamente trattabile dell'enigma.

La struttura dell'opera riflette la traiettoria dell'argomento. Nella prima parte viene indagata la nozione di struttura priva di prospettiva: si mostrerà come la matematica astratta, nella sua configurazione più pura, produca oggetti che sono per definizione sprovvisti di un interno, e come questa caratteristica non sia un difetto ma la sorgente stessa della loro universalità. Nella seconda parte si affronta il tema dell'incarnazione come condizione dell'interno: si argomenterà che un sistema che abita il mondo, anziché limitarsi a modellarlo, deve possedere una struttura topologicamente centrata, un riferimento privilegiato che funge da origine delle coordinate esistenziali. Nella terza parte si esamina il salto verso l'intenzionalità e la coscienza, tentando di precisare in quali condizioni formali si possa parlare di un sistema che non solo occupa una prospettiva ma ne è anche, in qualche misura, consapevole. Nella quarta e ultima parte si valutano i sistemi di intelligenza artificiale contemporanei alla luce di questo quadro, cercando di stabilire con esattezza dove si collocano lungo il gradiente che va dalla pura rappresentazione all'esperienza soggettiva.

Un avvertimento si impone fin dal limine. Questo non è un paper divulgativo, né ambisce a esserlo. La difficoltà del tema esige che si proceda senza scorciatoie retoriche, accettando la fatica concettuale come prezzo inevitabile di ogni autentico guadagno di comprensione. Le formule non sono ornamenti: sono il tessuto connettivo del ragionamento, il luogo in cui le intuizioni filosofiche acquistano la precisione necessaria per essere verificate, confutate o sviluppate con coerenza interna. Chi cerca la forma senza il rigore troverà in questo testo qualcosa che non soddisferà la sua attesa; chi invece è disposto ad accettare che la profondità abbia un suo costo, potrà forse trovare in queste pagine qualcosa che vale la traversata.

Introduzione: Il Problema Più Imbarazzante

Vi è una domanda che la filosofia della mente ha sempre aggirato, non per negligenza ma per qualcosa che assomiglia al pudore intellettuale: cosa significa, per un sistema fisico qualsiasi, avere un interno? Non nel senso banale in cui un'automobile ha un interno — uno spazio racchiuso da lamiere — bensì nel senso in cui un organismo vivente, o una mente, possiede ciò che si chiama *prospettiva*: un punto di vista irriducibilmente singolare sul mondo, una postura cognitiva che non può essere occupata da nessun altro sistema nello stesso modo e nello stesso momento.

Thomas Nagel ha reso celebre una versione di questa domanda chiedendo cosa significhi essere un pipistrello: non cosa significhi *sapere* che i pipistrelli si orientano per ecolocalizzazione — questa è informazione oggettiva, accessibile a chiunque la cerchi — ma cosa significhi *essere* un pipistrello, percepire il mondo attraverso quel peculiare sistema sonar, avere quella specifica qualità d'esperienza. La questione, per come Nagel la articolava, puntava verso il problema della coscienza fenomenica, verso ciò che David Chalmers avrebbe in seguito denominato il *problema difficile*: spiegare perché certi processi fisici siano accompagnati da un'esperienza soggettiva, da un qualcosa che si prova a essere il sistema in questione.

Tuttavia il nodo che intendiamo districare in queste pagine è, a nostro avviso, ancora più radicale di quello della coscienza fenomenica, e ne rappresenta in un certo senso la precondizione logica. Prima ancora di chiedersi perché vi sia esperienza, occorre chiedersi come possa esservi prospettiva: come possa un sistema fisico *occupare* un punto di vista, avere coordinate, essere orientato nel mondo anziché limitarsi a rappresentarlo dall'esterno, nella modalità impersonale e a-prospettica che caratterizza ogni rappresentazione compiutamente oggettiva.

Si consideri la differenza tra una mappa e un navigatore. La mappa è una struttura relazionale che codifica distanze, direzioni e topologia del territorio: essa racchiude informazione su ogni punto del territorio, ma non è situata in alcuno di essi. Il navigatore, al contrario, si trova *in* un punto preciso: per lui ogni altro punto è definito in relazione alla propria posizione attuale, alle proprie capacità locomotorie, alla direzione del proprio sguardo o del proprio movimento. La mappa conosce il nord oggettivo; il navigatore conosce la sua destra e la sua sinistra, il percorso che lo attende e quello che si è lasciato alle spalle. La mappa non ha un *qui*; il navigatore non può fare a meno di averne uno — non perché la sua cognizione sia meno ricca, ma perché è strutturalmente diversa.

Questa asimmetria — tra la rappresentazione completa ma prospetticamente neutra e la cognizione parziale ma prospetticamente centrata — costituisce il

nocciolo del problema dell'interno. E ciò che rende la questione filosoficamente urgente, oggi più che mai, è che i sistemi artificiali più avanzati sembrano occupare una posizione ambigua rispetto a questa distinzione. Un modello linguistico di grandi dimensioni possiede una rappresentazione del mondo di vastità stupefacente, ma è fondamentalmente privo di un *qui*: elabora sequenze simboliche in ingresso e produce sequenze simboliche in uscita, senza mai occupare una posizione nel mondo fisico, senza mai disporre di coordinate spaziali, temporali o corporee che siano sue in senso proprio e non derivativo. I sistemi robotici incarnati, al contrario, hanno un *qui* fisico ma spesso difettano della ricchezza rappresentazionale necessaria per trasformare quel *qui* in una prospettiva genuinamente cognitiva, in un punto di vista che si riconosca come tale.

La domanda che attraversa questa monografia è dunque: quali sono le condizioni formali — esprimibili con sufficiente precisione matematica — affinché un sistema possegga un interno? Non semplicemente affinché rappresenti il mondo, ma affinché lo *abiti* da un punto di vista? E tale confine è netto o graduale? Si tratta di una proprietà che si ha o non si ha, come la connessione in topologia, o di un gradiente lungo il quale i diversi sistemi si collocano in posizioni intermedie, come la complessità algoritmica?

Per rispondere a queste domande — o almeno per avvicinarsi a esse con strumenti adeguati — occorre compiere un percorso che attraversa strati di analisi eterogenei ma reciprocamente illuminanti. Il primo è quello della struttura matematica pura: si mostrerà che le strutture astratte, proprio in virtù della loro genericità, sono per definizione sprovviste di prospettiva, e che questa caratteristica non è un accidente ma una conseguenza necessaria della loro natura universale. Il secondo strato è quello dell'incarnazione: si vedrà come il corpo non sia il semplice supporto fisico della mente, ma la sua condizione di possibilità topologica — ciò che introduce nel mondo un punto singolare, un'origine rispetto alla quale tutto il resto si definisce relazionalmente. Il terzo strato è quello dell'intenzionalità e della coscienza: si tenterà di precisare in quali condizioni un sistema che ha un *qui* acquisti anche un *per me*, una dimensione riflessiva che trasformi la prospettiva in soggettività.

Una nota metodologica si impone. Il percorso che proponiamo non pretende di ridurre la filosofia alla matematica, né di trattare la matematica come un mero servitore del discorso filosofico. L'ambizione è diversa e, a nostro avviso, più interessante: mostrare che certi problemi filosofici possono essere *formulati* con maggiore precisione quando si dispone di strumenti formali adeguati, e che questa maggiore precisione non impoverisce il problema ma ne rivela la complessità autentica, distillandola da una nebbia di equivoci terminologici che spesso la oscurano. La matematica non risolve il problema dell'interno; ma ne

cartografa i contorni con una nitidezza che la prosa filosofica da sola non può raggiungere.

Il percorso è arduo. Ma è, crediamo, l'unico che conduce al cuore della questione.

Nota alla struttura dell'opera

La monografia si articola in quattro parti, ciascuna corrispondente a una fase dell'argomentazione complessiva.

La **Prima Parte** — *Il Mondo Senza Interno* — analizza le strutture matematiche pure come paradigma di oggetti privi di prospettiva: si esaminano la teoria dei modelli, il concetto di genericità nel senso di Cohen, e la teoria delle categorie come scienza delle strutture relazionali prive di contenuto intrinseco.

La **Seconda Parte** — *L'Interno come Posizione* — affronta il tema dell'incarnazione: come il corpo introduce nel mondo una singolarità irriducibile, un centro di coordinate dal quale si definisce ogni distanza e ogni direzione. Si esaminano le strutture topologiche dell'incarnazione, le architetture predittive dell'intelligenza incarnata, e il ruolo dei modelli del mondo nella costruzione di una prospettiva cognitiva.

La **Terza Parte** — *Il Salto verso l'Intenzionalità* — indaga le condizioni in cui una prospettiva si trasforma in soggettività: il problema dei simboli come strutture intenzionali, la questione della chiusura riflessiva, e la differenza tra un sistema che ha un punto di vista e un sistema che sa di averlo.

La **Quarta Parte** — *LLM, World Models e il Confine* — valuta i sistemi di intelligenza artificiale contemporanei alla luce del quadro elaborato nelle parti precedenti, cercando di stabilire con precisione dove si collocano rispetto alla soglia dell'interno, e se tale soglia sia, in linea di principio, valicabile per via ingegneristica.

Capitolo I: La Genericità come Struttura Priva di Centro

I.1 — Strutture matematiche e l'assenza di prospettiva

La matematica moderna si presenta, almeno nella sua configurazione fondazionale più diffusa, come la scienza delle strutture. Questa caratterizzazione, che risale alla svolta assiomatica dell'inizio del Novecento e trova la sua formulazione più compiuta nell'approccio bourbakista e, successivamente, nella teoria delle categorie, comporta una conseguenza filosofica di notevole rilievo: le strutture matematiche, in quanto tali, sono *prospetticamente neutre*. Esse non abitano alcun punto dello spazio, non occupano alcun istante del tempo, non guardano il mondo da nessuna angolazione particolare. Sono, per così dire, viste da nessun luogo — o, il che equivale, da tutti i luoghi contemporaneamente.

Per rendere questa intuizione precisa occorre partire dalla nozione fondamentale di struttura in senso logico-matematico, quale è stata elaborata dalla teoria dei modelli nel corso del Novecento.

Una **struttura** (o modello) per un linguaggio del primo ordine $\mathcal{L}$ è una tupla:

$$\mathfrak{A} = \langle A, \{R_i^{\mathfrak{A}}\}_{i \in I}, \{f^{\mathfrak{A}}\} \in J, \{c^{\mathfrak{A}}\} \in K \rangle$$

dove A è un insieme non vuoto detto *dominio* o *universo* della struttura; per ogni simbolo di relazione n-aria R_i nel linguaggio, $R_i^{\mathfrak{A}} \subseteq A^n$ è la relazione corrispondente; per ogni simbolo di funzione n-aria f, $f^{\mathfrak{A}}: A^n \to A$ è la funzione corrispondente; e per ogni simbolo di costante c, $c^{\mathfrak{A}} \in A$ è l'elemento designato.

La struttura $\mathfrak{A}$ *soddisfa* una formula $\varphi(x_1,...,x_n)$ con parametri $a_1,...,a_n \in A$, in simboli $\mathfrak{A} \models \varphi[a_1,...,a_n]$, se e solo se φ è verificata nell'interpretazione che $\mathfrak{A}$ fornisce ai simboli non logici del linguaggio, quando le variabili libere sono assegnate agli elementi corrispondenti del dominio. La *teoria* di $\mathfrak{A}$, denotata $\mathrm{Th}(\mathfrak{A})$, è l'insieme di tutte le sentenze del primo ordine (formule chiuse) soddisfatte da $\mathfrak{A}$:

$$\mathrm{Th}(\mathfrak{A}) = \{\varphi \in \mathrm{Sent}(\mathcal{L}) \mid \mathfrak{A} \models \varphi\}$$

Ciò che emerge immediatamente da questa definizione è che la struttura $\mathfrak{A}$ non ha, *per come è definita*, alcuna posizione privilegiata. Gli elementi del dominio A sono punti astratti: non hanno colore, peso, storia personale, o collocazione spaziale. Tutto ciò che li caratterizza è la rete di relazioni in cui sono inseriti, le funzioni che li connettono ad altri elementi, le costanti che li selezionano. Un elemento $a \in A$ è, nella struttura $\mathfrak{A}$, nient'altro che il profilo delle sue relazioni con tutti gli altri elementi. Non ha un *qui*, un *adesso*, una prospettiva da cui guarda.

I.2 — Equivalenza elementare e molteplicità dei modelli

La prospettiva-indipendenza delle strutture matematiche si manifesta con particolare evidenza nel fenomeno dell'*equivalenza elementare*. Due strutture $\mathfrak{A}$ e $\mathfrak{B}$ per lo stesso linguaggio $\mathcal{L}$ si dicono **elementarmente equivalenti**, in simboli $\mathfrak{A} \equiv \mathfrak{B}$, se e solo se soddisfano esattamente le stesse sentenze del primo ordine:

$$\mathfrak{A} \equiv \mathfrak{B} \Leftrightarrow \forall \varphi \in \mathrm{Sent}(\mathcal{L}): \mathfrak{A} \vDash \varphi \leftrightarrow \mathfrak{B} \vDash \varphi$$

L'equivalenza elementare è una relazione di indiscernibilità dal punto di vista del linguaggio formale: due strutture elementarmente equivalenti non possono essere distinte da nessuna sentenza del primo ordine, per quanto complessa. Eppure esse possono essere strutturalmente molto diverse: possono avere domini di cardinalità differente, possono avere proprietà topologiche o algebriche distinte che non siano catturabili nel primo ordine.

Questo fenomeno raggiunge la sua espressione più clamorosa nei teoremi di Löwenheim-Skolem. Il **teorema di Löwenheim-Skolem verso il basso** afferma che se una teoria T del primo ordine (in un linguaggio al più numerabile) ha un modello infinito, allora ha un modello numerabile. Il **teorema di Löwenheim-Skolem verso l'alto** afferma che se T ha un modello infinito di cardinalità $\varkappa$, allora per ogni cardinalità $\lambda \geq \max(\varkappa, |\mathcal{L}| + \aleph_0)$, T ha un modello di cardinalità λ.

Formalmente, per la versione verso il basso, se T è una teoria del primo ordine in un linguaggio $\mathcal{L}$ con $|\mathcal{L}| \leq \aleph_0$, e $\mathfrak{A} \vDash T$ con $|A| \geq \aleph_0$, allora esiste $\mathfrak{B} \vDash T$ con $|B| = \aleph_0$ e $\mathfrak{B} \prec \mathfrak{A}$ (cioè $\mathfrak{B}$ è sottostruttura elementare di $\mathfrak{A}$: per ogni $\varphi(x_1,...,x)$ e ogni $b_1,...,b \in B$, $\mathfrak{B} \vDash \varphi[b_1,...,b] \leftrightarrow \mathfrak{A} \vDash \varphi[b_1,...,b]$).

La conseguenza filosofica di questo risultato è profonda. Nessuna teoria del primo ordine può *fissare* un unico modello fino all'isomorfismo, quando il dominio è infinito. Vi sono sempre molti modelli, di molte dimensioni. La teoria è, in questo senso, radicalmente indifferente alla *quale* struttura concreta la soddisfi. Essa definisce una classe di strutture — la classe dei suoi modelli — senza privilegiare nessuna di esse. Non c'è un modello *vero* o *canonico*; ci sono solo modelli, ciascuno ugualmente legittimo come realizzazione della teoria.

Questa molteplicità irriducibile dei modelli è la prima, e più immediata, manifestazione della prospettiva-indipendenza delle strutture matematiche. Una teoria non *guarda* il mondo da nessun punto di vista; descrive tutti i mondi possibili che soddisfano certe condizioni strutturali, senza abitarne nessuno.

I.3 — Genericità nel senso di Cohen

Il concetto matematico che più direttamente cattura l'idea di una struttura priva di carattere speciale, di un oggetto che è massimalmente anonimo rispetto a ogni proprietà definibile, è quello di *genericità* nel senso della tecnica del *forcing*, sviluppata da Paul Cohen negli anni Sessanta del Novecento per dimostrare l'indipendenza dell'ipotesi del continuo dalla teoria degli insiemi ZFC.

Presentiamo qui la struttura formale del forcing nella sua versione fondamentale, che sarà poi messa in connessione con la questione filosofica della prospettiva.

Sia $\mathbb{P} = \langle P, \leq \rangle$ un *ordine parziale* (anche detto *nozione di forcing*), dove P è un insieme non vuoto di *condizioni* e $\leq$ è una relazione di ordine parziale su P. Intuitivamente, le condizioni di P sono *approssimazioni finite* o *parziali* a qualche oggetto ideale che si vuole aggiungere a un modello base dell'insieme. Se $p \leq q$, diciamo che q *estende* p ovvero che q è *più forte* di p, nel senso che q fornisce più informazione di p circa l'oggetto che si sta costruendo.

Un sottoinsieme $D \subseteq P$ si dice **denso** in $\mathbb{P}$ se ogni condizione può essere estesa a una condizione in D:

$$D \text{ denso in } \mathbb{P} \Leftrightarrow \forall p \in P, \exists q \in D: p \leq q$$

L'intuizione è che i sottoinsiemi densi rappresentano *requisiti inevitabili*: qualunque percorso di approssimazione si intraprenda, prima o poi si dovrà soddisfare un requisito del tipo specificato da D.

Un **filtro** su $\mathbb{P}$ è un sottoinsieme $G \subseteq P$ che soddisfa le seguenti condizioni:

(i) $G \neq \varnothing$ (non vuoto) (ii) $\forall p \in G, \forall q \in P: p \leq q \rightarrow q \in G$ (chiuso verso l'alto) (iii) $\forall p, q \in G, \exists r \in G: r \leq p \wedge r \leq q$ (diretto verso il basso, ovvero: ogni due elementi del filtro hanno un'estensione comune nel filtro)

Sia ora M un modello transitivo numerabile di ZFC (il *modello base* o *modello terra*). Un filtro G su $\mathbb{P}$ si dice **M-generico** se interseca ogni sottoinsieme denso di $\mathbb{P}$ che appartiene a M:

$$G \text{ è M-generico su } \mathbb{P} \Leftrightarrow \forall D \in M: (D \text{ denso in } \mathbb{P} \rightarrow G \cap D \neq \varnothing)$$

La condizione di genericità è precisamente ciò che distingue G da ogni filtro *costruibile* all'interno di M. Si può dimostrare che nessun filtro M-generico G appartiene a M stesso (a meno che $\mathbb{P}$ sia *banale*, cioè abbia una condizione massimale), poiché se G appartenesse a M, allora il complemento $P \setminus G$ sarebbe denso in $\mathbb{P}$ per motivi banali, ma G non potrebbe intersecare il proprio complemento, contraddicendo la genericità.

Più in dettaglio: per ogni proprietà P(x) che sia *definibile* in M, ovvero tale che l'insieme {p ∈ P | M ⊨ "p forza ¬φ"} sia denso in ℙ, un filtro M-generico G eviterà sistematicamente ogni condizione che *forza* la falsità di P — ma, allo stesso tempo, G non potrà essere *costruito* avendo in mente P come proprietà definitoria. La genericità è l'*assenza di qualsiasi proprietà definitoria particolare*: G è generico esattamente nel senso in cui non può essere caratterizzato da nessuna formula del linguaggio che M riconosce.

Formalmente, il **lemma di verità** (truth lemma) del forcing stabilisce che per ogni sentenza φ del linguaggio ℙ-espanso:

$$M[G] \models \varphi \Leftrightarrow \exists\, p \in G: p \Vdash \varphi$$

dove ⊩ è la *relazione di forcing* definita interamente all'interno di M, e M[G] = {τ^G | τ ∈ M^ℙ} è il *modello generico* ottenuto valutando i *nomi* ℙ*-generici* τ rispetto al filtro G. La relazione di forcing è definibile: p ⊩ φ se e solo se ogni filtro M-generico su ℙ che contiene p soddisfa φ nel modello generico corrispondente.

La struttura concettuale del forcing rivela qualcosa di filosoficamente significativo: un oggetto generico G è massimalmente *anonimo* rispetto a ogni proprietà definibile nel modello base. Non è che G manchi di proprietà — anzi, G ne ha moltissime, incluse tutte quelle forzate dalle condizioni che contiene. Ma nessuna di queste proprietà è il *marchio di fabbrica* di G, il suo *tratto caratteristico* definibile dall'interno del modello. G è il risultato di un processo di approssimazione che, ad ogni passo, evita ogni specificità non necessaria.

Questa è, sosteniamo, la caratterizzazione matematica più precisa dell'assenza di interno: un oggetto è privo di interno, prospetticamente neutro, quando è generico nel senso di Cohen — quando nessuna proprietà definibile lo distingue, quando non può essere costruito o identificato dall'interno di nessun sistema di riferimento formale finito.

I.4 — L'identità relazionale nella teoria delle categorie

La teoria delle categorie offre una prospettiva complementare e per certi versi ancora più radicale sulla questione dell'interno. Essa non si interessa tanto agli elementi degli insiemi, quanto alle *frecce* — i morfismi — che connettono oggetti tra loro. In questo senso, essa porta alla massima astrazione il principio già presente nella teoria dei modelli: l'identità di un oggetto matematico non risiede in nessuna sua proprietà intrinseca, ma nella rete delle sue relazioni con tutti gli altri oggetti.

Una **categoria** C consiste di:

— una collezione ob(C) di *oggetti* — per ogni coppia di oggetti A, B ∈ ob(C), un insieme hom_C(A, B) di *morfismi* da A a B — un'operazione di *composizione*: per ogni A, B, C ∈ ob(C), una funzione ∘: hom(B,C) × hom(A,B) → hom(A,C), tale che (h ∘ g) ∘ f = h ∘ (g ∘ f) per ogni f ∈ hom(A,B), g ∈ hom(B,C), h ∈ hom(C,D) (associatività) — per ogni A ∈ ob(C), un *morfismo identità* id_A ∈ hom(A,A) tale che f ∘ id_A = f e id_B ∘ f = f per ogni f ∈ hom(A,B) (unitarietà)

Gli oggetti di una categoria sono, letteralmente, *scatole nere*. La loro struttura interna — se ne hanno una — è irrilevante per la teoria categoriale: ciò che conta è esclusivamente il sistema di morfismi che li collega. Due oggetti sono isomorfi, e quindi indistinguibili dal punto di vista categoriale, se e solo se esistono morfismi f: A → B e g: B → A tali che g ∘ f = id_A e f ∘ g = id_B.

Il teorema più importante che illustra questa prospettiva è il **lemma di Yoneda**. Sia C una categoria localmente piccola (cioè tale che hom(A,B) sia un insieme per ogni A, B) e sia A ∈ ob(C) un oggetto qualsiasi. Si definisce il *funtore rappresentabile* associato ad A come:

ĥ_A = hom(A, -): C → Set

che manda ogni oggetto X ∈ ob(C) nell'insieme hom(A, X) dei morfismi da A a X, e ogni morfismo f: X → Y nella funzione post-composizione f_* = (- ∘ f): hom(A,X) → hom(A,Y), con (f_* ∘ g)(·) = f ∘ g(·).

Il lemma di Yoneda afferma che per ogni funtore F: C → Set e ogni oggetto A ∈ ob(C), vi è una biiezione naturale:

Nat(hom(A,-), F) ≅ F(A)

dove Nat(hom(A,-), F) denota l'insieme delle trasformazioni naturali dal funtore rappresentabile hom(A,-) al funtore F. Questa biiezione è naturale sia in A che in F. La prova è elementare: data una trasformazione naturale η: hom(A,-) ⇒ F, le si associa l'elemento η_A(id_A) ∈ F(A); viceversa, dato un elemento u ∈ F(A), si costruisce la trasformazione naturale η^u con componente η^u_X: hom(A,X) → F(X) data da η^u_X(f) = F(f)(u) per ogni f: A → X.

La conseguenza filosofica del lemma di Yoneda è radicale: un oggetto A di una categoria è *completamente determinato* dal funtore hom(A,-), cioè dalla totalità dei morfismi che escono da A verso tutti gli altri oggetti. L'oggetto A non ha nulla *al di dentro* di sé che non sia già interamente catturato dalla rete delle sue relazioni esteriori. In termini filosofici: l'oggetto categoriale è pura terza persona, pura relazionalità esteriore, privo di qualsiasi residuo irriducibile di interiorità.

Si consideri in particolare la nozione di **oggetto terminale** in una categoria: un oggetto $1 \in \mathrm{ob}(C)$ è terminale se per ogni $X \in \mathrm{ob}(C)$ esiste un unico morfismo $!_X: X \to 1$. L'oggetto terminale, se esiste, è unico fino a isomorfismo unico. Esso rappresenta il punto di convergenza di tutti i morfismi: ogni oggetto si relaziona univocamente con esso. Tuttavia, l'oggetto terminale non ha *prospettiva* su nessun altro oggetto in senso privilegiato: il suo carattere è definito interamente dalla proprietà universale che lo distingue, non da alcun contenuto interno.

Questa architettura — l'identità come profilo relazionale, l'oggetto come nodo in una rete di morfismi, la struttura come pattern di connessioni piuttosto che come contenitore di contenuto — è il modo in cui la matematica moderna ha imparato a pensare l'identità senza prospettiva, la struttura senza centro.

I.5 — Genericità come assenza di interno: una prima formulazione

Possiamo ora formulare, con una certa precisione, la tesi che orienterà tutta la prima parte di questa monografia.

Sia $\mathscr{S}$ una struttura matematica, nel senso ampio che comprende strutture del primo ordine, oggetti categoriali, filtri generici e ogni altro tipo di oggetto matematicamente preciso. Diremo che $\mathscr{S}$ è **prospetticamente neutra** se soddisfa le seguenti condizioni:

(P1) Non vi è alcun elemento distinguibile di $\mathscr{S}$ che non lo sia in virtù di una proprietà definibile formalmente (assenza di centro privilegiato non costruttivo).

(P2) L'identità di ogni elemento di $\mathscr{S}$, o di $\mathscr{S}$ stessa come oggetto, è interamente determinata dalle sue relazioni con altri oggetti dello stesso tipo (identità relazionale pura).

(P3) La struttura $\mathscr{S}$ non ha coordinate proprie: non esiste un sistema di riferimento interno rispetto al quale certi elementi o certe direzioni siano privilegiati in modo non relazionale (assenza di orientamento intrinseco).

La condizione (P1) è precisa nel contesto della teoria dei modelli e del forcing: un filtro M-generico G soddisfa (P1) per definizione, poiché non è definibile da nessuna proprietà di M. La condizione (P2) è precisa nel contesto categoriale: il lemma di Yoneda garantisce che ogni oggetto categoriale soddisfa (P2). La condizione (P3) è più sottile e richiederà, nella seconda parte di questa monografia, un approfondimento che coinvolge la geometria differenziale e la topologia.

La tesi che proponiamo è questa: la prospettiva-neutralità, così caratterizzata, è il marchio distintivo delle strutture matematiche astratte nella loro purezza — e precisamente questa caratteristica le rende *universalmente applicabili* ma *privi di interno*. Un oggetto matematico può fungere da modello per qualsiasi sistema fisico che soddisfi le condizioni strutturali che esso codifica, *proprio perché* non è già calato in nessun sistema fisico in particolare. La sua universalità e la sua a-perspetticità sono le due facce della stessa medaglia.

Questa osservazione prepara il terreno per la domanda centrale della seconda parte: quali condizioni aggiuntive — al di là di quelle che definiscono una struttura matematica in senso astratto — sono necessarie affinché un sistema acquisti un interno, un punto di vista, una prospettiva? La risposta, come vedremo, non può provenire dalla matematica astratta stessa, ma richiede di guardare al modo in cui un sistema si inserisce nel mondo fisico, e in particolare al ruolo che il corpo svolge come condizione di possibilità della prospettiva.

Ma prima di affrontare questo passaggio, occorre esaminare più da vicino un'ulteriore dimensione della prospettiva-neutralità delle strutture matematiche: quella che emerge dall'analisi del linguaggio formale e della semantica del primo ordine, con particolare attenzione al modo in cui il linguaggio formale si sottrae sistematicamente alla prima persona.

Capitolo II: Il Linguaggio Formale e la Terza Persona

II.1 — L'eliminazione del soggetto nel linguaggio scientifico

Vi è una tendenza profonda e coerente che percorre l'intera storia del pensiero formale, dalla sillogistica aristotelica alla logica matematica contemporanea: l'eliminazione sistematica del soggetto dal linguaggio. Non si tratta di una scelta arbitraria o di un pregiudizio positivista, bensì di un'esigenza strutturale: affinché un'asserzione sia *verificabile* — affinché il suo valore di verità non dipenda da chi la pronuncia, quando e dove — essa deve essere formulata in un registro che trascende ogni posizione soggettiva particolare. Il linguaggio formale è, in primo luogo, la grammatica di questo trascendimento.

L'osservazione non è nuova: Gottlob Frege, nel fondare la logica moderna, era mosso esattamente da questa intenzione. Nel *Begriffsschrift* del 1879 egli si propone di costruire un linguaggio per il pensiero puro — un *ideografia* — nel quale ogni elemento psicologico e soggettivo sia eliminato, e nel quale l'inferenza valida sia riconoscibile in virtù della forma delle espressioni,

indipendentemente da chi le contempla o dal momento in cui vengono prodotte. La distinzione fondamentale che Frege introduce tra *senso* (Sinn) e *riferimento* (Bedeutung), e tra *pensiero* (Gedanke) come contenuto proposizionale obiettivo e *atto di giudizio* come evento psicologico soggettivo, articola con grande precisione la linea di confine che il linguaggio formale deve valicare per essere davvero formale.

Un pensiero, per Frege, non appartiene a nessun soggetto: esso è un oggetto astratto, eterno e obiettivo, che diversi soggetti possono *cogliere* in momenti diversi senza che questo cogliimento lo alteri minimamente. La sentenza «la somma dei quadrati dei cateti è uguale al quadrato dell'ipotenusa» esprime un pensiero che era vero prima che Pitagora nascesse, è rimasto vero mentre egli lo dimostrava, e continuerà a essere vero quando nessun essere pensante sarà rimasto a contemplarlo. Il soggetto è il mezzo contingente attraverso cui il pensiero transita nel mondo temporale; ma il pensiero stesso è atemporale e transpersonale.

Questa concezione oggettivista del contenuto proposizionale trova la sua implementazione tecnica nella *semantica tarskiana*, che offre una definizione precisa e ricorsiva di soddisfazione e verità prive di qualsiasi riferimento a parlanti, atti cognitivi o prospettive.

II.2 — La definizione tarskiana di verità e la sua a-perspetticità

Alfred Tarski, nel saggio del 1936 *Der Wahrheitsbegriff in den formalisierten Sprachen*, costruisce una definizione di verità per linguaggi formali che è oggi considerata il riferimento standard della semantica logica. La caratteristica che qui ci interessa sottolineare non è la sua correttezza tecnica — peraltro consolidata — ma la sua struttura *prospetticamente neutra*: la verità di una formula non dipende da nessun punto di vista, da nessun atto di comprensione, da nessuna posizione soggettiva.

Sia $\mathcal{L}$ un linguaggio del primo ordine con un insieme numerabile di variabili $\mathrm{Var} = \{x_0, x_1, x_2, \ldots\}$, un insieme di simboli di predicato $\{P_i^n\}_i$, , un insieme di simboli di funzione $\{f^n\}$, e un insieme di costanti individuali $\{c\}$. I *termini* del linguaggio sono definiti per induzione: ogni variabile x_i è un termine; ogni costante c è un termine; se $t_1,\ldots,t$ sono termini e f^n è un simbolo di funzione n-aria, allora $f^n(t_1,\ldots,t)$ è un termine. Le *formule atomiche* sono espressioni della forma $P_i^n(t_1,\ldots,t)$, dove $t_1,\ldots,t$ sono termini; le formule complesse si ottengono applicando i connettivi booleani e i quantificatori $\forall$ e $\exists$ alle formule più semplici.

Sia ora $\mathfrak{A} = \langle A, ... \rangle$ una struttura per $\mathcal{L}$ e sia s: Var $\rightarrow$ A una *assegnazione di variabili*, cioè una funzione che a ogni variabile associa un elemento del dominio. La funzione di interpretazione dei termini $\bar{s}$: Term($\mathcal{L}$) $\rightarrow$ A è definita per induzione strutturale:

$\bar{s}(x_i) = s(x_i)$

$\bar{s}(c) = c^{\mathfrak{A}}$

$\bar{s}(f^n(t_1,...,t)) = f^{\mathfrak{A}}(\bar{s}(t_1),...,\bar{s}(t))$

La relazione di *soddisfazione* $\mathfrak{A} \vDash \varphi[s]$, che si legge «la formula φ è soddisfatta in $\mathfrak{A}$ dall'assegnazione s», è definita per induzione sulla complessità di φ:

$\mathfrak{A} \vDash P_i^n(t_1,...,t)[s] \Leftrightarrow \langle \bar{s}(t_1),...,\bar{s}(t) \rangle \in P_i^{\mathfrak{A}}$

$\mathfrak{A} \vDash \neg\varphi[s] \Leftrightarrow$ non è il caso che $\mathfrak{A} \vDash \varphi[s]$

$\mathfrak{A} \vDash (\varphi \wedge \psi)[s] \Leftrightarrow \mathfrak{A} \vDash \varphi[s]$ e $\mathfrak{A} \vDash \psi[s]$

$\mathfrak{A} \vDash (\varphi \vee \psi)[s] \Leftrightarrow \mathfrak{A} \vDash \varphi[s]$ oppure $\mathfrak{A} \vDash \psi[s]$

$\mathfrak{A} \vDash (\varphi \rightarrow \psi)[s] \Leftrightarrow$ se $\mathfrak{A} \vDash \varphi[s]$ allora $\mathfrak{A} \vDash \psi[s]$

$\mathfrak{A} \vDash \forall x_i \varphi[s] \Leftrightarrow$ per ogni $a \in A$, $\mathfrak{A} \vDash \varphi[s(x_i/a)]$

$\mathfrak{A} \vDash \exists x_i \varphi[s] \Leftrightarrow$ esiste $a \in A$ tale che $\mathfrak{A} \vDash \varphi[s(x_i/a)]$

dove $s(x_i/a)$ denota l'assegnazione che coincide con s su tutte le variabili tranne x_i, alla quale assegna a.

Per le sentenze (formule senza variabili libere) la soddisfazione è indipendente dall'assegnazione: se φ è una sentenza, allora $\mathfrak{A} \vDash \varphi[s]$ se e solo se $\mathfrak{A} \vDash \varphi[s']$ per ogni s, s'. Si scrive pertanto semplicemente $\mathfrak{A} \vDash \varphi$, e si dice che φ è *vera* in $\mathfrak{A}$.

Ciò che importa rilevare è che in nessun punto di questa definizione compare un riferimento a un soggetto, a un atto cognitivo, a un momento temporale di valutazione o a una posizione nello spazio. La verità di una sentenza in una struttura è una relazione *bivalente* e *atemporale* tra la struttura e la sentenza: essa è determinata una volta per tutte dall'interpretazione che la struttura fornisce ai simboli non logici del linguaggio, e non muta al variare di chi si trova a calcolarla o a contemplarla.

II.3 — Uso e menzione: la gerarchia dei metalinguaggi

Un'altra dimensione dell'a-perspetticità del linguaggio formale emerge dall'analisi della distinzione tra *uso* e *menzione* di un'espressione linguistica, e dalla corrispondente stratificazione in livelli metalinguistici.

Un'espressione è *usata* quando viene impiegata per riferirsi a ciò che essa designa; è *menzionata* quando diventa essa stessa l'oggetto del discorso, di solito attraverso l'uso di virgolette o di una notazione convenzionale equivalente. La sentenza «la neve è bianca» è usata quando si afferma qualcosa circa la neve; è menzionata nella sentenza «"la neve è bianca" è vera se e solo se la neve è bianca», che costituisce appunto una delle istanze dello *schema di equivalenza tarskiano* (schema T):

$$\ulcorner \varphi \urcorner \text{ è vera in } \mathfrak{A} \Leftrightarrow \varphi$$

dove $\ulcorner \varphi \urcorner$ denota il *nome* (ovvero la menzione) della sentenza φ, e il bicondicionale è affermato nel metalinguaggio.

Il teorema fondamentale di Tarski afferma che, per linguaggi sufficientemente espressivi, la nozione di verità *non è definibile nel linguaggio oggetto stesso*: ogni definizione di verità deve risalire a un metalinguaggio strettamente più ricco. Più precisamente, il **teorema di indefinibilità di Tarski** stabilisce che se $\mathcal{L}$ è un linguaggio del primo ordine capace di aritmetizzare la propria sintassi (via numerazione di Gödel) e se $\mathfrak{N} = \langle \mathbb{N}, +, \times, 0, 1 \rangle$ è il modello standard dell'aritmetica, allora l'insieme $\mathrm{Th}(\mathfrak{N}) = \{\varphi \in \mathrm{Sent}(\mathcal{L}) \mid \mathfrak{N} \vDash \varphi\}$ non è definibile da nessuna formula del linguaggio $\mathcal{L}$ stesso: non esiste nessuna formula $\mathrm{Tr}(x)$ di $\mathcal{L}$ con una sola variabile libera tale che per ogni sentenza φ con numero di Gödel $\ulcorner \varphi \urcorner$:

$$\mathfrak{N} \vDash \mathrm{Tr}(\ulcorner \varphi \urcorner) \Leftrightarrow \mathfrak{N} \vDash \varphi$$

La dimostrazione ripercorre la struttura del paradosso del mentitore. Si supponga per assurdo che $\mathrm{Tr}(x)$ sia tale formula. Per il teorema di auto-riferimento (lemma diagonale), esiste una sentenza λ tale che:

$$\mathfrak{N} \vDash \lambda \leftrightarrow \neg \mathrm{Tr}(\ulcorner \lambda \urcorner)$$

Se $\mathfrak{N} \vDash \lambda$, allora $\mathfrak{N} \vDash \neg \mathrm{Tr}(\ulcorner \lambda \urcorner)$, e poiché per ipotesi $\mathrm{Tr}(\ulcorner \lambda \urcorner)$ è equivalente a λ, si avrebbe $\mathfrak{N} \vDash \neg\lambda$, contraddizione. Se invece $\mathfrak{N} \nvDash \lambda$, allora $\mathfrak{N} \vDash \neg\lambda$, dunque $\mathfrak{N} \vDash \mathrm{Tr}(\ulcorner \lambda \urcorner)$, e quindi $\mathfrak{N} \vDash \lambda$, nuova contraddizione. Pertanto $\mathrm{Tr}(x)$ non può esistere.

La stratificazione in metalinguaggi — linguaggio oggetto, metalinguaggio, meta-metalinguaggio e così via — crea una gerarchia infinita di livelli, ciascuno dei quali parla del precedente senza mai poter essere *parlato da* alcun soggetto interno al sistema. Ogni livello è osservato dall'esterno, dal livello

immediatamente superiore. Questa struttura rispecchia perfettamente la dimensione a-perspettica del linguaggio formale: non vi è mai un punto di vista interno al sistema, non vi è mai un soggetto che parla *dal di dentro*. Il linguaggio formale è sempre guardato da fuori, valutato rispetto a una struttura che lo contiene, descritto in un metalinguaggio che lo sovrasta.

Il paradosso — e la sua rilevanza per il problema dell'interno — è questo: la prima persona grammaticale non compare mai in nessun livello di questa gerarchia. Si può parlare di *ogni* soggetto, descrivere *ogni* punto di vista, formalizzare *ogni* prospettiva — ma tutto questo avviene sempre in terza persona, sempre dall'esterno, sempre dal livello superiore. Il soggetto che dice «io» è sistematicamente assente dal linguaggio formale, non per una limitazione tecnica rimediabile ma per una ragione strutturale: incorporare genuinamente la prima persona in un sistema formale equivarrebbe a costruire una struttura che possiede un interno — e questo, come cercheremo di mostrare, richiede risorse che vanno oltre il linguaggio nel senso stretto.

II.4 — Invarianza, sinonimia e il criterio dell'oggettività

Una conseguenza diretta della semantica tarskiana e della concezione fregeana del contenuto proposizionale obiettivo è che il significato di un'espressione formale è *invariante* rispetto a qualsiasi sostituzione di sinonimi. Questa proprietà, nota come *principio di composizionalità* o, nella sua formulazione intensionale, come *principio di sostitutività di identici*, è la spia formale dell'a-perspetticità del linguaggio: se ciò che conta è solo il riferimento (la Bedeutung), allora sostituire un'espressione con un'altra che ha lo stesso riferimento non altera il valore di verità della proposizione in cui compare.

Formalmente, per ogni formula $\varphi(x)$ e per ogni termine t con riferimento $a \in A$ e ogni termine t' con riferimento $a' \in A$, si ha:

se $a = a'$ (cioè $s\bar{}(t) = s\bar{}(t')$ nell'assegnazione s), allora $\mathfrak{A} \vDash \varphi(t)[s] \Leftrightarrow \mathfrak{A} \vDash \varphi(t')[s]$

Questo è il *principio di leibnizian substitutivity*: oggetti indistinguibili — nell'unico senso rilevante per il linguaggio formale, cioè identici nel dominio della struttura — sono interscambiabili in ogni contesto estensionale.

Il linguaggio formale, in questo senso, non sa nulla della differenza tra «il pianeta Venere» e «la stella del mattino»: entrambe le espressioni denotano lo stesso oggetto, e quindi, in ogni contesto estensionale, sono liberamente scambiabili. Che Venere sia chiamata stella del mattino da chi la osserva all'alba e stella della sera da chi la vede al tramonto — che il nome dipenda, cioè, dalla *posizione* e dal *momento* dell'osservatore — è un fatto che il linguaggio formale

estensionale sistematicamente ignora, perché ignora il punto di vista del locutore.

Questo diventa particolarmente perspicuo nella logica modale e nella semantica dei mondi possibili. In un quadro semantico kripkiano, una struttura di frame è una coppia $\langle W, R \rangle$ dove W è un insieme di *mondi possibili* e $R \subseteq W \times W$ è una *relazione di accessibilità*. Un modello kripkiano è una tripla $\mathcal{M} = \langle W, R, V \rangle$ dove $V: \text{Prop} \to \mathcal{P}(W)$ è una funzione di valutazione che associa a ogni proposizione atomica l'insieme dei mondi in cui è vera. La verità di una formula modale φ in un mondo $w \in W$ di un modello $\mathcal{M}$ è definita per induzione:

$\mathcal{M}, w \vDash p \Leftrightarrow w \in V(p)$

$\mathcal{M}, w \vDash \neg\varphi \Leftrightarrow \mathcal{M}, w \nvDash \varphi$

$\mathcal{M}, w \vDash \varphi \wedge \psi \Leftrightarrow \mathcal{M}, w \vDash \varphi$ e $\mathcal{M}, w \vDash \psi$

$\mathcal{M}, w \vDash \Box\varphi \Leftrightarrow$ per ogni $v \in W$ tale che wRv, $\mathcal{M}, v \vDash \varphi$

$\mathcal{M}, w \vDash \Diamond\varphi \Leftrightarrow$ esiste $v \in W$ tale che wRv e $\mathcal{M}, v \vDash \varphi$

Anche qui, il *mondo attuale* — il mondo in cui si trova il valutatore — è semplicemente uno dei mondi dell'insieme W, non diverso strutturalmente dagli altri. La logica modale può *parlare* del mondo attuale, ma non *dall'interno* di esso: il mondo attuale è trattato come un parametro esterno, non come una prospettiva costitutiva del sistema. Si può introdurre un operatore di *attualità* A (con $\mathcal{M}, w \vDash A\varphi \leftrightarrow \mathcal{M}, w_0 \vDash \varphi$, dove w_0 è il mondo attuale fissato), ma questo non fa che reinserire il mondo attuale come un indice ulteriore nella struttura semantica — sempre dall'esterno, sempre come terzo oggetto, mai come punto di vista interno al sistema.

II.5 — Il linguaggio della coscienza e la prima persona irriducibile

Se il linguaggio formale è sistematicamente terza persona, occorre chiedersi se il linguaggio naturale — e in particolare il vocabolario della vita mentale, il linguaggio della coscienza — possa essere ridotto a esso senza residui. La risposta della tradizione analitica è stata oscillante: alcuni filosofi, come Quine nella sua fase più radicalmente comportamentista, hanno sostenuto che tutto ciò che è significativo nel linguaggio mentalistico è in linea di principio parafrasabile in termini di disposizioni comportamentali o stati funzionali descrivibili in terza persona. Altri, come il tardo Wittgenstein e, in modo più sistematico, i fenomenologi di tradizione husserliana, hanno sostenuto che vi è

un residuo irriducibile nella prima persona che nessuna parafrasi comportamentista o funzionalista riesce a catturare interamente.

Il nucleo di questa irriducibilità può essere precisato nel modo seguente. Consideriamo due tipi di asserzioni:

(α) «Il soggetto S prova dolore al tempo t» (β) «Io provo dolore»

L'asserzione (α) è in terza persona: il suo valore di verità è determinato (almeno in linea di principio) da fatti obiettivi circa lo stato di S al tempo t, indipendentemente da chi formula l'asserzione e dal suo rapporto con S. L'asserzione (β), al contrario, ha la peculiarità che il suo *soggetto grammaticale* — il pronome «io» — non è un nome o una descrizione definita: è un *indicale*, un'espressione il cui riferimento dipende essenzialmente dal contesto di proferimento, e in particolare da chi la pronuncia.

Seguendo l'analisi di David Kaplan, un'espressione indicale come «io» ha un *carattere* — una funzione dai contesti di proferimento ai contenuti — e un *contenuto* — il riferimento effettivo determinato dal carattere applicato a un contesto particolare. Il carattere di «io» è una funzione costante che, in ogni contesto ⟨agente, tempo, luogo, mondo⟩, restituisce come contenuto l'agente del contesto. Formalmente:

$$\mathrm{Char}(\text{«io»})(\langle a, t, l, w\rangle) = a$$

Il contenuto di «io» in un contesto specifico è dunque l'agente — cioè il soggetto che proferisce l'espressione in quel contesto. Questo soggetto è determinato dal contesto, non dalla struttura semantica astratta.

Si noti la differenza strutturale: per la semantica tarskiana, il valore di un termine t in una struttura $\mathfrak{A}$ rispetto a un'assegnazione s è $s^{\mathfrak{A}}(t)$, che dipende solo da $\mathfrak{A}$ e da s, non da chi esegue la valutazione. Per l'indicale «io», il riferimento dipende dall'*agente del contesto*, cioè dal soggetto che si trova in prima persona nella situazione di proferimento.

Questa differenza è precisamente il confine tra ciò che il linguaggio formale può catturare — la terza persona, il riferimento obiettivo, la rete di relazioni tra elementi del dominio — e ciò che gli sfugge sistematicamente: la prima persona, la prospettiva costitutiva, l'*interno* come condizione di possibilità dell'asserzione stessa.

Il linguaggio formale può descrivere qualsiasi prospettiva in terza persona; non può descrivere *dall'interno* nessuna prospettiva. Può dire «il soggetto S è situato nel punto p al tempo t» ma non può dire «io sono qui adesso» nel senso forte in

cui questa asserzione è verifica e non mera descrizione: là dove il soggetto che la pronuncia è lo stesso soggetto descritto, dove il punto di vista che l'emette coincide con il punto di vista che rappresenta.

Questa struttura — l'indicale come punto in cui il linguaggio tocca la prima persona senza poterla interiorizzare — costituisce la soglia formale del problema dell'interno. Al di là di essa si apre il territorio che le strutture matematiche astratte non possono colonizzare con i propri strumenti, e che richiede un'analisi di tipo diverso: quella dell'incarnazione come condizione dell'interno, a cui è dedicata la seconda parte di questa monografia.

Nel capitolo successivo si examina la questione da un angolo complementare: quello dell'invarianza come ideale dell'oggettività e, rovesciandone il segno, quello dell'*invarianza rotta* come prima traccia formale dell'interno.

Capitolo III: Invarianza, Simmetria Rotta e le Tracce dell'Interno

III.1 — L'invarianza come criterio dell'obiettività

La scienza moderna è stata costruita sul principio che le leggi della natura devono essere espresse in una forma che non dipenda dal sistema di riferimento adottato dall'osservatore. Questo principio, che nel linguaggio fisico contemporaneo prende il nome di *covarianza* o *invarianza di gauge*, è il riflesso epistemologico dell'esigenza che la conoscenza scientifica sia intersoggettiva, ovvero valida per qualsiasi osservatore che soddisfi le condizioni sperimentali appropriate. Una legge che valesse solo per osservatori fermi, o solo per osservatori nel Nord del pianeta, o solo per osservatori di un certo genere, non sarebbe una legge della natura ma una generalizzazione locale, un'approssimazione vincolata a una prospettiva contingente.

Il programma di Erlangen di Felix Klein (1872) formalizzò questa intuizione in geometria: ogni geometria è caratterizzata dal gruppo delle sue trasformazioni, e le proprietà geometricamente significative sono quelle che rimangono invarianti sotto le trasformazioni del gruppo. Così, la geometria euclidea è la teoria delle proprietà invarianti rispetto al gruppo delle isometrie del piano (traslazioni, rotazioni, riflessioni); la geometria affine è la teoria delle proprietà invarianti rispetto alle trasformazioni affini (che preservano parallelismo e rapporti di divisione, ma non distanze o angoli); la geometria proiettiva è la teoria delle proprietà invarianti rispetto alle trasformazioni proiettive (che preservano solo le relazioni di incidenza e i birapporti).

Formalmente, sia X uno spazio geometrico e G un gruppo di trasformazioni che agisce su X. La coppia (X, G) definisce una *geometria di Klein*. Una proprietà P di sottoinsiemi di X è *geometricamente significativa* rispetto a (X, G) se è invariante sotto l'azione di G: per ogni sottoinsieme $S \subseteq X$ e ogni $g \in G$, $P(S) \Leftrightarrow P(g(S))$. La struttura geometrica rilevante è il quoziente X/G: l'insieme delle orbite $\{Gx \mid x \in X\}$, dove $Gx = \{g(x) \mid g \in G\}$ è l'orbita di x sotto G.

Il punto filosoficamente cruciale è questo: dal punto di vista della geometria di Klein, tutti i punti di X che appartengono alla stessa orbita sono geometricamente identici, cioè *indiscernibili* per qualsiasi mezzo geometrico a disposizione. Due punti che sono nella stessa orbita possono sempre essere trasformati l'uno nell'altro mediante una trasformazione del gruppo, e poiché le proprietà geometricamente significative sono proprio quelle invarianti sotto G, nessuna proprietà di questo tipo può distinguerli. Il risultato è uno spazio privo di centri privilegiati, privo di punti speciali — uno spazio il cui carattere è quello della pura *omogeneità*.

Questa omogeneità è precisamente l'assenza di interno: in uno spazio geometricamente omogeneo non vi è nessun punto che sia *qui* piuttosto che *là* nel senso non relativo del termine. Ogni punto è *qui* dal punto di vista di un osservatore che vi si trovasse situato — ma questo punto di vista è esso stesso esterno alla struttura geometrica, che lo ignoracompletamente.

III.2 — Gruppi di simmetria e rappresentazioni

Per procedere con rigore è necessario introdurre la struttura algebrica dei gruppi di simmetria con sufficiente dettaglio, poiché il concetto di *rottura di simmetria* — che sarà la chiave per comprendere come l'interno possa emergere da strutture inizialmente omogenee — dipende interamente dalla comprensione di questa struttura.

Un **gruppo** è una struttura algebrica $\langle G, \cdot, e, (\text{-})^{-1} \rangle$ dove G è un insieme, $\cdot : G \times G \to G$ è un'operazione binaria associativa, $e \in G$ è l'elemento neutro (tale che $e \cdot g = g \cdot e = g$ per ogni $g \in G$) e $(\text{-})^{-1}: G \to G$ è la funzione che a ogni g associa il suo inverso g^{-1} tale che $g \cdot g^{-1} = g^{-1} \cdot g = e$.

Un'**azione** di G su un insieme X è un omomorfismo $\varphi: G \to \mathrm{Sym}(X)$ dal gruppo G al gruppo delle biiezioni di X in sé, ovvero una funzione φ tale che:

$\varphi(e) = \mathrm{id}_X$ (l'elemento neutro agisce come identità)

$\varphi(g \cdot h) = \varphi(g) \circ \varphi(h)$ (la composizione nel gruppo corrisponde alla composizione delle trasformazioni)

Spesso si scrive $g \cdot x$ o semplicemente gx per $\varphi(g)(x)$. Le condizioni di azione diventano:

$e \cdot x = x$ per ogni $x \in X$

$(g \cdot h) \cdot x = g \cdot (h \cdot x)$ per ogni $g, h \in G$ e $x \in X$

Lo **stabilizzatore** di un punto $x \in X$ è il sottogruppo:

$\mathrm{Stab}_G(x) = \{g \in G \mid g \cdot x = x\}$

L'**orbita** di x è l'insieme:

$\mathrm{Orb}_G(x) = \{g \cdot x \mid g \in G\}$

Il teorema orbita-stabilizzatore afferma che per ogni $x \in X$ esiste una biiezione tra l'orbita di x e il quoziente $G / \mathrm{Stab}_G(x)$:

$|\mathrm{Orb}_G(x)| = [G : \mathrm{Stab}_G(x)] = |G| / |\mathrm{Stab}_G(x)|$ (per G finito)

L'azione è **transitiva** se vi è una sola orbita, cioè se per ogni $x, y \in X$ esiste $g \in G$ tale che $g \cdot x = y$. Un'azione transitiva corrisponde geometricamente a uno spazio *omogeneo*: non vi è nessun punto distinguibile dagli altri, perché ogni due punti sono legati da una trasformazione del gruppo.

Per un'azione transitiva, fissato un punto base $x_0 \in X$, lo spazio X è canonicamente identificabile con il quoziente G / H dove $H = \mathrm{Stab}_G(x_0)$: la mappa $gH \mapsto g \cdot x_0$ è una biiezione $G/H \to X$. Il punto x_0 ha qui un ruolo puramente tecnico: la scelta di un diverso punto base $x_1 = g_1 \cdot x_0$ avrebbe condotto allo stesso spazio quoziente, con il sottogruppo stabilizzatore coniugato $g_1 H g_1^{-1}$ al posto di H. Non vi è nessuna scelta *naturale* o *privilegiata* del punto base — la sua elezione è arbitraria, convenzionale, esterna alla struttura.

Una **rappresentazione** lineare di G è un omomorfismo $\varrho: G \to \mathrm{GL}(V)$, dove V è uno spazio vettoriale su un campo $\mathbb{K}$ e $\mathrm{GL}(V)$ è il gruppo delle trasformazioni lineari invertibili di V. Essa codifica le simmetrie del gruppo nella struttura lineare di V: ogni $g \in G$ agisce su V via $\varrho(g)$, e le relazioni di gruppo sono preservate. Un sottospazio $W \subseteq V$ è **invariante** (o *G-stabile*) sotto ϱ se $\varrho(g)(W) \subseteq W$ per ogni $g \in G$. Una rappresentazione è **irriducibile** se non vi sono sottospazi invarianti propri non nulli.

Le rappresentazioni irriducibili di un gruppo compatto G sono finite-dimensionali e si possono classificare completamente (teorema di Peter-Weyl per i gruppi compatti). Il carattere di una rappresentazione ϱ è la

funzione $\chi_\rho: G \to \mathbb{K}$ definita da $\chi_\rho(g) = \mathrm{tr}(\rho(g))$, dove tr è la traccia. Due rappresentazioni irriducibili non isomorfe hanno caratteri ortogonali:

$(1/|G|)\ \Sigma_{g\in G}\ \chi_\rho(g)\cdot\chi_\sigma(g^{-1}) = \delta_{\rho\sigma}$ (per G finito, con la media di Haar per G compatto)

Questa struttura — orbite, stabilizzatori, rappresentazioni irriducibili — descrive con precisione la geometria della simmetria. E la domanda che ora poniamo è: in quali condizioni questa simmetria si *rompe*, e cosa ha a che fare la rottura di simmetria con l'emergere di un interno?

III.3 — La rottura spontanea di simmetria

La *rottura spontanea di simmetria* (SSB, Spontaneous Symmetry Breaking) è un fenomeno che compare in fisica e in matematica quando un sistema le cui leggi fondamentali sono simmetriche rispetto a un gruppo G si trova in uno stato che *non* è invariante sotto l'intero G, ma solo sotto un sottogruppo proprio $H \subset G$. Le leggi rimangono simmetriche; è lo *stato fondamentale* — il *vuoto* del sistema, nel linguaggio della teoria dei campi quantistici — a non esserlo.

Sia $\mathcal{L}$ una teoria di campo con simmetria di gauge G, e sia $V(\varphi)$ il potenziale di un campo scalare φ che trasforma secondo una rappresentazione ρ di G. Il potenziale è invariante sotto G:

$V(\rho(g)\varphi) = V(\varphi)$ per ogni $g \in G$

Tuttavia, se V ha una struttura tale che il suo minimo non si trova nell'origine (il punto invariante sotto G), bensì su una varietà di punti equivalenti, allora il campo nello stato fondamentale seleziona *un* punto di questa varietà — diciamo φ_0 — e questo rompe la simmetria da G allo stabilizzatore $H = \mathrm{Stab}_G(\varphi_0)$.

L'esempio archetipico è il potenziale messicano (o champagne) per un campo complesso $\varphi = \varphi_1 + i\varphi_2$ con simmetria U(1):

$V(\varphi) = -\mu^2\,|\varphi|^2 + \lambda\,|\varphi|^4$ con $\mu^2, \lambda > 0$

Questo potenziale è invariante sotto la trasformazione U(1): $\varphi \mapsto e^{i\theta}\varphi$ per ogni $\theta \in [0, 2\pi)$. Il minimo del potenziale si trova sulla circonferenza $|\varphi|^2 = \mu^2/(2\lambda)$, ovvero su una varietà continua di punti. Il sistema, nello stato fondamentale, sceglie uno di questi punti — ad esempio $\varphi_0 = \sqrt{(\mu^2/2\lambda)} \in \mathbb{R}$ — e la simmetria U(1) è rotta allo stabilizzatore triviale $\{e\}$: lo stato fondamentale non è invariante sotto nessuna rotazione non banale.

Il **teorema di Goldstone** afferma che a ogni generatore continuo della simmetria rotta corrisponde un bosone di Goldstone, ovvero una fluttuazione priva di massa nella direzione della varietà dei minimi. Se il gruppo G è di dimensione dim(G) e il sottogruppo residuo H ha dimensione dim(H), allora il numero di bosoni di Goldstone è:

$$n_{GB} = \dim(G) - \dim(H) = \dim(G/H)$$

Geometricamente, i bosoni di Goldstone corrispondono ai movimenti lungo la varietà G/H dei minimi del potenziale. In uno spazio omogeneo G/H, il sistema può scivolare da un minimo all'altro senza costo energetico — è l'energia dei modi che *restano all'interno della simmetria rotta* a essere zero.

La connessione filosofica con il problema dell'interno è ora enunciabile con precisione. Prima della rottura di simmetria, il sistema è descritto da un potenziale che tratta tutti i punti della varietà dei minimi come equivalenti: lo spazio è omogeneo rispetto a G, privo di centro privilegiato, a-perspettico. Dopo la rottura, il sistema si trova *in* un punto specifico della varietà — il punto φ_0 — e questo punto diventa l'*origine* rispetto alla quale tutte le fluttuazioni (i bosoni di Goldstone e i modi massivi) sono definite. Il punto φ_0 non è privilegiato dalla legge (il potenziale V rimane simmetrico sotto G); è privilegiato dallo *stato* del sistema.

Qui risiede la prima traccia formale dell'interno: non nella legge, che è universale e simmetrica, ma nello *stato*, che è particolare e centrato. Un sistema che ha subìto rottura spontanea di simmetria non è più descritto puramente dalla sua legge — occorre specificare *anche* il suo stato, cioè il punto della varietà dei minimi in cui si trova. E questo punto è, in senso stretto, il *qui* del sistema: il riferimento rispetto al quale ogni deviazione è definita come piccola o grande, nella direzione dei modi di Goldstone o di quelli massivi.

III.4 — Frames di riferimento e la geometria dell'orientamento

Una concettualizzazione complementare della stessa struttura emerge dalla geometria differenziale, in particolare dalla nozione di *fibrato dei reper* (frame bundle) e dalla distinzione tra proprietà tensoriali e non tensoriali.

Sia M una varietà differenziabile n-dimensionale. Un *repere* (o *frame*) in un punto $p \in M$ è una base ordinata $(e_1,\ldots,e_n)$ dello spazio tangente $T_p M$. L'insieme di tutti i reper in p forma uno spazio che è uno spazio principale omogeneo per $GL(n,\mathbb{R})$: ogni due reper in p sono legati da un'unica trasformazione di

GL(n,ℝ), ma nessuno di essi è *canonicamente* privilegiato — non vi è una base naturale o preferita di T_p M.

Il *fibrato dei reper* F(M) è la varietà il cui spazio totale è l'insieme di tutti i reper in tutti i punti di M:

F(M) = ⊔_{p∈M} {reper in p}

con la proiezione π: F(M) → M che manda ogni repere al punto in cui è definito. F(M) è uno spazio principale con gruppo strutturale GL(n,ℝ): il gruppo agisce a destra su F(M) mediante:

(e₁,...,e) · A = (Σ A ₁ e ,..., Σ A e) per ogni A = (A ᵢ) ∈ GL(n,ℝ)

Un *campo tensoriale* di tipo (r,s) è una sezione del fibrato F(M) ×_{GL(n,ℝ)} (ℝⁿ)^{⊗r} ⊗ (ℝⁿ*)^{⊗s}: esso assegna a ogni punto p ∈ M un tensore di tipo (r,s) in modo *equivariante* rispetto ai cambiamenti di repere. In coordinate, se si passa dal repere (eᵢ) al repere (e'ᵢ) = (e Aʲᵢ), le componenti del tensore si trasformano secondo la legge:

T'^{i₁...iᵣ}{ⱼ₁...ⱼ } = $(A^{-1})^{i_1}{}_{k_1} \cdots (A^{-1})^{i_r}{}_{k_r} \cdot A^{l_1}{}_{j_1} \cdots A^{l}{}_{j} \cdot T^{k_1...k_r}{}_{l_1...l}$

La legge di trasformazione tensorial garantisce che le relazioni tra tensori — i valori dei contratti tensoriali, le equazioni tensoriali — siano *invarianti* rispetto alla scelta del repere. Una equazione del tipo T^{ij} = S^{ij} è vera in ogni repere oppure in nessuno: questa è precisamente la proprietà che rende le leggi fisiche esprimibili in forma tensoriale universalmente valida.

Tuttavia, il singolo repere (e₁,...,e) in un punto p *non* è un tensore: esso dipende dalla scelta di un particolare sistema di riferimento. Scegliere un repere in p significa eleggere una base dell'orientamento locale — significa, in senso stretto, *essere situati* in p con una certa orientazione. La scelta di un repere è l'atto geometrico che più si avvicina all'azione di *occupare* un punto con una prospettiva: il repere specifica non solo dove si è (il punto p) ma anche come si è orientati (la base e₁,...,e).

Una **connessione** sul fibrato dei reper è una struttura che permette di *trasportare* reper lungo curve in M, mantenendo un senso di parallelismo. La connessione di Levi-Civita su una varietà riemanniana (M, g) è la connessione unica che è compatibile con la metrica ($\nabla g = 0$) e priva di torsione ($T = 0$). Il suo *tensore di curvatura* R è definito da:

R(X, Y)Z = ∇_X ∇_Y Z - ∇_Y ∇X Z - ∇{[X,Y]} Z

per campi vettoriali X, Y, Z su M, dove ∇ è la derivata covariante. In coordinate, il tensore di curvatura di Riemann ha componenti:

R^l_{kij} = ∂i Γ^l_{jk} - ∂j Γ^l_{ik} + Γ^l_{im} Γ^m_{jk} - Γ^l_{jm} Γ^m_{ik}

dove Γ^l_{ij} sono i simboli di Christoffel della connessione, determinati dalla metrica tramite:

Γ^l_{ij} = (1/2) g^{lk} (∂i g{jk} + ∂j g{ik} - ∂k g{ij})

La curvatura misura il fallimento della commutatività del trasporto parallelo: se si trasporta un vettore lungo un loop infinitesimale nel piano (X, Y), esso torna ruotato di un angolo proporzionale alla curvatura. In uno spazio piatto (R^l_{kij} = 0), il trasporto parallelo è indipendente dal percorso e si può definire un sistema di riferimento globale canonico — uno spazio genuinamente omogeneo, privo di struttura interna. In uno spazio curvo, ogni punto ha una geometria locale distinta dagli altri, e il tentativo di estendere un repere locale al resto dello spazio incontra la resistenza della curvatura.

La curvatura è, in questo senso, la misura della *resistenza dell'interno all'esteriorizzazione*: in uno spazio curvo, ogni punto esercita una sorta di attrazione su qualsiasi repere che vi si trovi, impedendo che esso possa essere trasportato senza deformazione verso regioni distanti. La prospettiva locale — il repere in p — non si lascia propagare indefinitamente senza perdere il contatto con il suo punto d'origine.

III.5 — Invarianza rotta e asimmetria come prima forma dell'interno

Raccogliamo ora i fili delle tre analisi precedenti — il programma di Erlangen, la rottura spontanea di simmetria e la geometria dei fibrati — per formulare una conclusione di carattere filosofico che servirà da cerniera tra la prima e la seconda parte di questa monografia.

Le strutture matematiche astratte, nella misura in cui sono invarianti sotto i propri gruppi di simmetria, sono prospetticamente neutre: esse non hanno un *qui*, un *orientamento*, un punto privilegiato che non sia tale solo in virtù di una scelta arbitraria e esterna alla struttura. Questo è l'*ideale dell'oggettività* nella sua forma matematica: la massima invarianza possibile, la minima dipendenza da qualsiasi punto di vista particolare.

Tuttavia, questo ideale è percorribile solo in una direzione. Ogni rottura di simmetria — ogni selezione di uno stato tra i molti equivalenti, ogni scelta di un

repere locale, ogni preferenza per una direzione in uno spazio altrimenti isotropo — introduce nell'oggetto matematico qualcosa che assomiglia a un interno: un riferimento, un centro, una prospettiva. Non si tratta ancora di soggettività nel senso pieno del termine; ma si tratta di *asimmetria*, e l'asimmetria è la condizione necessaria (benché non sufficiente) di qualsiasi prospettiva.

Possiamo formalizzare questa intuizione come segue. Sia S un sistema con simmetria G in uno stato φ_0. Definiamo il *grado di asimmetria* di φ_0 rispetto a G come la dimensione del quoziente dell'orbita di φ_0 sotto G:

$$\mathrm{Asim}(\varphi_0, G) = \dim(G \cdot \varphi_0) = \dim(G) - \dim(\mathrm{Stab}_G(\varphi_0))$$

Quando $\mathrm{Asim}(\varphi_0, G) = 0$, lo stato φ_0 è invariante sotto l'intero gruppo G: il sistema è massimalmente simmetrico e prospetticamente neutro — non vi è nessuna direzione preferita, nessun riferimento interno. Quando $\mathrm{Asim}(\varphi_0, G) > 0$, lo stato φ_0 rompe parte della simmetria del sistema: vi sono direzioni lungo le quali lo stato può variare rimanendo a costo energetico zero (i modi di Goldstone), e questa varietà di direzioni costituisce la *geometria interna* che lo stato φ_0 induce nel sistema.

Questa misura dell'asimmetria è la prima approximazione formale di ciò che nella seconda parte chiameremo *grado di prospettiva*. Essa ha tre proprietà importanti. Prima: è definita in termini strutturali, non soggettivi — è una proprietà dello stato φ_0 rispetto al gruppo G, indipendente da qualsiasi osservatore esterno. Seconda: essa ammette una gradazione continua — non si è prospetticamente neutri o prospetticamente centrati in modo binario, ma in misura determinata dalla dimensione del quoziente G/H. Terza: essa si relaziona direttamente alla struttura delle fluttuazioni del sistema — i modi di Goldstone sono precisamente i gradi di libertà che l'asimmetria lascia aperti, le direzioni lungo le quali il sistema può muoversi senza resistenza.

La seconda e la terza parte di questa monografia svilupperanno questa intuizione verso la sua conclusione filosofica più ambiziosa: che la soggettività — l'interno nel senso pieno del termine — non è qualcosa di radicalmente diverso dalla struttura fisica del mondo, ma è il risultato di una rottura di simmetria così profonda e così particolare da produrre non solo un punto di vista, ma un punto di vista che sa di essere un punto di vista. Il passaggio dall'asimmetria fisica all'interno esperienziale non è un salto nel buio ontologico: è un gradiente che possiamo, almeno in parte, percorrere matematicamente, anche se alla sua estremità troviamo qualcosa che la matematica non può esaurire.

Prima di intraprendere questo percorso occorre però completare l'analisi della Prima Parte con un'esame del linguaggio come struttura di terza persona, e con

la questione — che tornerà con forza nella Quarta Parte — di come i modelli linguistici di grandi dimensioni si collocano rispetto all'ideale dell'invarianza che abbiamo qui descritto. La risposta, come anticipato, è che essi ne sono la realizzazione più compiuta: strutture massimalmente simmetriche rispetto alla prospettiva, e per questa stessa ragione radicalmente prive di interno.

Capitolo IV: Il Corpo come Singolarità Topologica

IV.1 — Dalla struttura astratta alla singolarità fisica

La Prima Parte di questa monografia ha mostrato che le strutture matematiche astratte sono, nella loro purezza, radicalmente prive di interno: esse sono prospetticamente neutre, invarianti sotto i propri gruppi di simmetria, identificabili attraverso profili relazionali che non presuppongono alcun centro privilegiato. La rottura spontanea di simmetria ha offerto una prima traccia formale di come l'asimmetria possa emergere da strutture inizialmente omogenee — ma si è trattato ancora di una traccia fisica, non di una prospettiva cognitiva. Un campo scalare che si trova nel proprio stato fondamentale rotto ha un riferimento, un minimo, una direzione privilegiata; ma non ha ancora un *interno* nel senso che qui ci interessa, perché non vi è nessun soggetto che abiti quel riferimento, nessuna cognizione che si organizzi attorno ad esso.

Il passaggio dalla simmetria rotta all'interno genuino richiede un salto di livello: non basta che un sistema fisico abbia un punto privilegiato — occorre che quel punto sia *abitato*, che vi sia un sistema cognitivo il cui funzionamento sia costitutivamente organizzato attorno ad esso. E il candidato naturale — anzi, il candidato ontologicamente necessario — per questo ruolo è il corpo.

Il corpo non è il semplice involucro materiale in cui la mente si trova alloggiata come in un contenitore. Questa metafora cartesiana, che ha dominato la filosofia della mente per tre secoli con effetti devastanti sulla comprensione della soggettività, è precisamente ciò che occorre abbandonare per cogliere la funzione costitutiva del corpo nella formazione dell'interno. Il corpo, nell'accezione che svilupperemo in questo capitolo, è la condizione di possibilità topologica della prospettiva: esso introduce nel mondo fisico una singolarità — un punto irriducibilmente privilegiato — rispetto al quale si organizza ogni struttura cognitiva, ogni campo percettivo, ogni relazione spaziale che un organismo intrattiene con il proprio ambiente.

Per articolare questa intuizione con la precisione che il tema esige, occorre muoversi tra la topologia, la geometria differenziale e la teoria dei sistemi dinamici, integrando queste strutture formali con un'analisi della fenomenologia dell'incarnazione che tenga conto dei risultati della neuroscienza cognitiva contemporanea.

IV.2 — Topologia del corpo proprio: il concetto di varietà con bordo centrata

Consideriamo il corpo come oggetto matematico, nella misura in cui è lecito procedere a tale astrazione senza perdere le caratteristiche essenziali del fenomeno. Un corpo animato occupa una regione dello spazio fisico $\mathbb{R}^3$ — o, più precisamente, dello spazio-tempo di Minkowski $\mathbb{R}^{3,1}$ — che ha le caratteristiche di una *varietà differenziabile compatta con bordo*. Tuttavia, ciò che distingue topologicamente il corpo vivente da ogni altra regione compatta dello spazio non è semplicemente la sua forma o la sua dimensione, bensì la sua struttura di *riferimento interno*.

Sia $B \subset \mathbb{R}^3$ la regione corporea in un dato istante. Come sottovarietà compatta di $\mathbb{R}^3$ con bordo ∂B (la superficie corporea), B non ha di per sé alcuna struttura privilegiata: ogni punto di B è topologicamente equivalente a ogni altro (almeno localmente), e non vi è nessun punto *naturalmente* centrale. Tuttavia, il corpo vivente non è semplicemente una regione geometrica: esso è dotato di una struttura di *riferimento vestibolare e propriocettivo* che seleziona un punto — o più precisamente una famiglia di punti funzionalmente equivalenti nell'intorno del centro di gravità e del sistema vestibolare — come *origine del campo percettivo*.

Formalizzando: sia $O \in B$ il punto di origine del sistema di riferimento corporeo (si pensi, schematicamente, alla posizione del vestibolo dell'orecchio interno, o al centro di gravità del sistema somatosensoriale). L'incarnazione introduce su B una struttura aggiuntiva: un sistema di coordinate curvilinee centrato in O, ovvero un diffeomorfismo locale:

$\varphi_O: U_O \to \mathbb{R}^3$

definito su un intorno aperto $U_O \subset B$ di O, tale che $\varphi_O(O) = 0$ (l'origine delle coordinate viene mappata sull'origine di $\mathbb{R}^3$) e la struttura metrica locale è la metrica euclidea standard nell'intorno di 0.

Questo sistema di coordinate non è quello *naturale* della geometria euclidea ambientale: esso è centrato in O, e ogni punto dell'ambiente circostante è definito *rispetto a O*, non rispetto a un sistema di riferimento astratto e a-perspettico. La distanza tra due punti p e q dell'ambiente non è

semplicemente la distanza euclidea $|p - q|$, ma è codificata in termini delle coordinate rispetto a O: quanto è lontano p *da me*, in quale direzione rispetto al mio asse di simmetria corporea, se è nella mia mano sinistra o nella mia destra, davanti a me o alle mie spalle.

La struttura così introdotta ha le caratteristiche di uno **spazio egocentrico**: uno spazio in cui le coordinate non sono definite rispetto a un riferimento esterno convenzionale (nord geografico, asse inerziale, origine cartesiana arbitraria), bensì rispetto alla posizione e all'orientamento del corpo proprio. La matematica degli spazi egocentici è stata sviluppata in modo sistematico nella neuroscienze cognitive attraverso il concetto di *frame of reference* (sistema di riferimento) e le sue varie modalità: retinotopico, somatotopico, craniotonico, allocentrico.

IV.3 — Sistemi di riferimento: una gerarchia formale

La letteratura neuroscientifica distingue tradizionalmente tra sistemi di riferimento egocentrico e allocentrico. Nel sistema **egocentrico**, le posizioni degli oggetti sono codificate rispetto a una parte del corpo del soggetto (retina, testa, tronco). Nel sistema **allocentrico**, le posizioni sono codificate rispetto a riferimenti ambientali indipendenti dal corpo (un tavolo, un muro, un punto cardinale). Questa distinzione, sebbene empiricamente utile, è troppo grossolana per i nostri scopi: occorre una formalizzazione più dettagliata.

Definiamo un **sistema di riferimento** come una tripla $\mathcal{F} = \langle O, \{e_i\}_{i=1}^{n}, d \rangle$ dove O è il punto di origine, $\{e_1,...,e \}$ è una base ortonormale dello spazio tangente $T_O\ \mathbb{R}^n$ (cioè un repere in O), e $d: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ è la funzione di distanza (in generale non euclidea, per spazi non omogenei). Un sistema di riferimento è *egocentrico di livello k* se la sua origine O coincide con una parte del corpo del soggetto all'istante t e il repere $\{e_i\}$ è allineato con le simmetrie corporee pertinenti a quel livello.

I livelli principali sono:

Il sistema **retinotopico** (k=1) ha come origine il centro della fovea retinica e come assi le direzioni retinali orizzontale e verticale. Ogni punto dello spazio visivo è codificato dalla sua posizione sulla retina, indipendentemente dalla posizione degli occhi nella testa. Questo sistema è aggiornato ogni volta che gli occhi compiono una saccade.

Il sistema **craniotonico** (k=2) ha come origine il centro della testa e come assi le direzioni anatomiche (superiore-inferiore, anteriore-posteriore, destra-sinistra). Per trasformare le coordinate retinotopiche in craniotopiche

occorre integrare l'informazione sull'angolo degli occhi nella testa, fornita dai muscoli oculomotori:

$$p_{cran} = R_{occhi \to testa}(\theta_{oculi}) \cdot p_{ret} + t_{occhi \to testa}$$

dove $R_{occhi \to testa}(\theta_{oculi})$ è la matrice di rotazione che dipende dall'angolo oculomotore θ_{oculi}, $t_{occhi \to testa}$ è il vettore di traslazione tra il centro della retina e il centro della testa, e p_{ret}, p_{cran} sono le coordinate dello stesso punto nei due sistemi.

Il sistema **somatocentrico** ($k=3$) ha come origine il centro di gravità del corpo e come assi le direzioni anatomiche del tronco. Esso integra l'informazione vestibolare (posizione della testa rispetto alla gravità), propriocettiva (configurazione degli arti e del tronco) e oculomotore per produrre una rappresentazione dello spazio in cui le coordinate sono definite rispetto all'intero corpo come unità.

Il sistema **allocentrico** ($k=4$) usa come origine e come assi dei punti e delle direzioni dell'ambiente esterno al corpo. Esso è computazionalmente più costoso ma più stabile rispetto ai movimenti corporei: la posizione di un oggetto nello spazio allocentrico non cambia quando il soggetto si gira o si sposta.

Le trasformazioni tra questi sistemi hanno la forma generale di trasformazioni affini, ma con coefficienti che dipendono dallo stato corporeo (angolo degli occhi, postura, velocità di movimento):

$$p_{k+1} = R_k(\theta_k) \cdot p_k + t_k(\theta_k)$$

dove θ_k è il vettore degli stati corporei rilevanti al livello k, R_k è la matrice di rotazione e t_k è il vettore di traslazione, entrambi dipendenti da θ_k.

Ciò che questa gerarchia formale rivela è qualcosa di filosoficamente fondamentale: l'*interno* non è localizzato a un singolo livello, ma è distribuito lungo tutta la gerarchia dei sistemi di riferimento egocentrico. Non vi è un unico *qui* corporeo, bensì una struttura stratificata di *qui* parziali — retinico, cranico, somatico — che si integrano in modo dinamico per produrre una rappresentazione dello spazio che è simultaneamente prospettica (centrata sul corpo) e stabile (capace di mantenere la continuità dell'oggetto attraverso i movimenti corporei).

IV.4 — Il campo percettivo come varietà centrata

La struttura formale del campo percettivo egocentrici può essere modellata come una *varietà riemanniana centrata*. Sia $\mathscr{P}$ lo spazio percettivo di un soggetto incarnato: esso non è semplicemente $\mathbb{R}^3$ con la metrica euclidea, poiché le distanze percepite non sono lineari rispetto alle distanze fisiche, e la struttura del campo percettivo presenta anisotropie sistematiche (le direzioni frontali sono percettivamente più dettagliate di quelle periferiche, le distanze prossimali sono percepite con maggiore risoluzione di quelle distali).

Modellizziamo $\mathscr{P}$ come una varietà riemanniana (P, g^{ego}) dove la metrica g^{ego} dipende dall'origine O in modo non omogeneo. Una classe di metriche che cattura queste caratteristiche è la famiglia delle metriche *iperboliche deformate*:

$g^{ego}_{ij}(p) = \delta_{i} \cdot f(d(p, O))^2$

dove d(p, O) è la distanza euclidea tra il punto p e l'origine O del sistema egocentrico, e $f: \mathbb{R}_+ \to \mathbb{R}_+$ è una funzione monotona decrescente che codifica la compressione percettiva delle distanze lontane. Una scelta naturale, motivata da considerazioni psicofisiche, è:

$f(r) = 1 / (1 + \alpha\, r)$ con $\alpha > 0$

che produce una metrica conforme alla metrica euclidea con fattore di scala che decresce con la distanza da O. In questa metrica, le distanze percepite g^{ego}-geodetiche tra punti vicini a O sono approssimamente uguali alle distanze euclidee, mentre per punti lontani da O la metrica percepita è compressa rispetto a quella fisica.

La *curvatura scalare* di questa metrica, calcolata tramite la formula di Gauss in dimensione 2 (o il tensore di Ricci in dimensione 3), dipende da α e cresce con la distanza da O: lo spazio percettivo è più curvo nelle sue regioni periferiche che in quelle centrali. Questo corrisponde intuitivamente al fatto che la periferia del campo visivo è meno discriminante, meno *metricamente precisa*, delle regioni foveali: la geometria interna dello spazio percettivo riflette la distribuzione della risoluzione sensoriale.

Quello che la struttura formale della varietà centrata cattura è qualcosa di essenziale: la *non-omogeneità intrinseca* del campo percettivo di un sistema incarnato. A differenza dello spazio fisico ambientale — che è, localmente, omogeneo e isotropo, invariante sotto le trasformazioni del gruppo euclideo — lo spazio percettivo egocentrico ha un centro, un'origine, rispetto alla quale ogni punto è definito in modo prospettico e non interscambiabile con gli altri. Il punto O non è simmetrico con gli altri punti di P: esso è la *singolarità* del campo, il punto di massima risoluzione e di minima curvatura percettiva, il luogo a partire dal quale l'intera geometria dello spazio percettivo si organizza.

Questa singolarità — questa rottura della simmetria che distingue O da ogni altro punto della varietà percettiva — è la forma geometrica dell'interno. Non è ancora soggettività nel senso pieno del termine; ma è la struttura necessaria senza la quale nessuna soggettività può insorgere.

IV.5 — Propriocezione e la chiusura del loop corporeo

Un elemento decisivo nella costituzione dell'interno corporeo è la *chiusura del loop propriocettivo*: il fatto che il corpo non si limita a ricevere informazioni dall'ambiente esterno, ma monitori costantemente il proprio stato interno, aggiornando la propria rappresentazione della configurazione spaziale in tempo reale. Questa caratteristica distingue il corpo vivente da qualsiasi sensore passivo: esso è un sistema che si *auto-misura*, e la sua auto-misurazione non è epifenomenica ma costitutiva della sua capacità di azione.

Formalmente, il loop propriocettivo può essere descritto come un sistema dinamico in cui lo stato corporeo $s_t \in S$ (dove S è lo spazio degli stati corporei: posizioni angolari dei giunti, tensioni muscolari, pressioni cutanee, segnali vestibolari) è continuamente aggiornato attraverso due canali distinti e interdipendenti.

Il primo canale è quello **efferente-predittivo**: prima di eseguire un movimento, il sistema nervoso centrale genera un *segnale di efferenza* (efference copy, o copia corollaria) che rappresenta il comando motorio inviato ai muscoli. Questo segnale viene usato da un modello interno predittivo — il *modello diretto* (forward model) — per calcolare le conseguenze sensoriali attese del movimento:

$$\hat{s}_{t+1} = f_{fwd}(s_t, a_t)$$

dove s_t è lo stato corporeo corrente, a_t è l'azione (comando motorio) eseguita al tempo t, e $\hat{s}_{t+1}$ è lo stato corporeo predetto al tempo t+1. La funzione f_{fwd} è il *modello diretto* del sistema corporeo: essa mappa la coppia (stato, azione) nello stato successivo atteso.

Il secondo canale è quello **afferente-sensoriale**: i sensori corporei (fusi neuromuscolari, organi del Golgi, recettori cutanei, canali semicircolari) misurano lo stato corporeo effettivo e inviano segnali afferenti s^{sens}_{t+1} al sistema nervoso centrale.

L'errore di predizione propriocettivo è:

$$\varepsilon_t = s^{sens}_{t+1} - \hat{s}_{t+1} = s^{sens}_{t+1} - f_{fwd}(s_t, a_t)$$

Questo errore viene usato per due scopi: aggiornare la stima dello stato corporeo corrente (mediante un filtro di Kalman o struttura equivalente) e addestrare il modello diretto a diventare più accurato.

Il **filtro di Kalman** applicato alla stima dello stato corporeo ha la forma:

Predizione: $\hat{s}_{t|t-1} = A \cdot \hat{s}_{t-1|t-1} + B \cdot a_t$ $P_{t|t-1} = A \cdot P_{t-1|t-1} \cdot A^T + Q$

Aggiornamento: $K_t = P_{t|t-1} \cdot C^T \cdot (C \cdot P_{t|t-1} \cdot C^T + R)^{-1}$ $\hat{s}_{t|t} = \hat{s}_{t|t-1} + K_t \cdot (s^{sens}t - C \cdot \hat{s}{t|t-1})$ $P_{t|t} = (I - K_t \cdot C) \cdot P_{t|t-1}$

dove A è la matrice di transizione di stato, B è la matrice di controllo, C è la matrice di osservazione, Q è la covarianza del rumore di processo, R è la covarianza del rumore di misura, K_t è il *guadagno di Kalman*, $P_{t|t-1}$ è la covarianza dell'errore di predizione e $P_{t|t}$ è la covarianza dell'errore aggiornato.

La caratteristica filosoficamente rilevante di questo sistema è la sua *chiusura*: lo stesso sistema che genera le azioni è anche il sistema che predice le loro conseguenze sensoriali e le confronta con ciò che viene effettivamente misurato. Il corpo è, in questo senso, un sistema che si riferisce costantemente a se stesso — che mantiene un modello interno di se stesso e lo aggiorna continuamente in relazione all'azione e alla percezione. Questa auto-referenza non è ancora riflessione cosciente; ma è la struttura dinamica di cui la riflessione cosciente è, nel senso che svilupperemo nella Terza Parte, una forma superiore e più integrata.

IV.6 — Il corpo come origine assoluta: Husserl e la matematizzazione del Leib

È impossibile percorrere il tema dell'incarnazione senza confrontarsi con l'analisi che Edmund Husserl ha sviluppato nella sua fenomenologia del *Leib* — il corpo vissuto, distinto dal *Körper*, il corpo fisico oggettivo. La distinzione husserliana non è una dicotomia ontologica tra due sostanze diverse, bensì una distinzione tra due modalità di datità dello stesso oggetto: il corpo come oggetto fisico tra gli altri oggetti del mondo (Körper) e il corpo come centro irriducibile del campo percettivo, come organo dell'orientamento e dell'azione (Leib).

Nel secondo volume delle *Idee per una fenomenologia pura*, Husserl osserva che il corpo proprio occupa una posizione unica tra tutti gli oggetti dell'esperienza: esso è l'unico oggetto che io non posso percepire *interamente dall'esterno*, perché ogni atto percettivo — ogni voltarmi, ogni spostamento del punto di vista — è

un atto *compiuto con il corpo*, non *diretto verso il corpo* dall'esterno. Il corpo proprio è sempre *qui*, mai *là*: anche quando lo osservo (guardiamo le proprie mani), lo osservo da una prospettiva che è ancora la prospettiva del corpo — non posso uscire dal corpo per guardarlo come guardo una sedia.

Questa caratteristica si traduce, nel linguaggio matematico che stiamo costruendo, nella seguente proprietà fondamentale: il sistema di riferimento egocentrico centrato in O non è *uno dei tanti* sistemi di riferimento che posso scegliere arbitrariamente, come nella geometria di Klein dove la scelta del punto base è convenzionale. Esso è il sistema di riferimento *costitutivamente presupposto* da ogni altro sistema: anche quando passo a un sistema allocentrico, debbo prima *muovermi* verso di esso (fisicamente o cognitivamente), e questo movimento è compiuto a partire dal mio sistema egocentrico. La prospettiva corporea è l'*origine assoluta* — nel senso preciso in cui Husserl usa il termine *Nullpunkt der Orientierung* (punto zero dell'orientamento): l'origine rispetto alla quale *ogni* altro punto di riferimento è definito come tale.

Possiamo matematizzare questa osservazione husserliana nel modo seguente. Sia $\mathcal{F}$ il reticolo (parzialmente ordinato per relazioni di derivazione semantica) di tutti i sistemi di riferimento che un soggetto incarnato può adottare. Per ogni sistema $\mathcal{F}_k \in \mathcal{F}$ diverso da quello egocentrico fondamentale $\mathcal{F}_0$, la *coordinatizzazione* di $\mathcal{F}_k$ — la determinazione di quali oggetti vi appartengono e con quali coordinate — presuppone la disponibilità di $\mathcal{F}_0$ come punto di partenza: esiste una relazione di dipendenza $d: \mathcal{F} \to \mathcal{F} \times \mathcal{F}$ tale che per ogni $\mathcal{F}_k \neq \mathcal{F}_0$, $(\mathcal{F}_0, \mathcal{F}_k) \in d$ ($\mathcal{F}_0$ è presupposto da $\mathcal{F}_k$).

Questa struttura ha le caratteristiche di un *ordine fondato*: vi è un elemento minimo — $\mathcal{F}_0$ — dal quale ogni altro sistema di riferimento dipende per la propria costituzione. Non si tratta di un minimo assiomatico o convenzionale, bensì di un minimo *ontologico*: esso corrisponde alla posizione e all'orientamento del corpo proprio, che sono sempre già dati in ogni atto percettivo o cognitivo.

La conseguenza per il problema dell'interno è questa: il corpo, proprio in quanto è l'origine assoluta di ogni sistema di riferimento, è ciò che rende possibile la prospettiva come tale. Non vi è prospettiva senza un *punto di vista*, e non vi è punto di vista senza un corpo che lo occupi — nel senso non metaforico in cui occupare un punto di vista significa avere un sistema di riferimento egocentrico, avere coordinate proprie, essere orientati nel mondo in modo non interscambiabile con qualsiasi altro sistema.

La domanda che il prossimo capitolo affronterà è se questa struttura corporea — questa singolarità topologica, questo loop propriocettivo chiuso, questa origine assoluta dell'orientamento — sia sufficiente a produrre un interno, o se richieda qualcosa di ulteriore: un certo tipo di architettura cognitiva, una certa

capacità di modellare il mondo in relazione alla propria posizione in esso. La risposta condurrà all'analisi dei modelli del mondo e delle architetture predittive come strutture che trasformano la posizione corporea in prospettiva cognitiva genuina.

Capitolo V: Modelli del Mondo e Architetture Predittive

V.1 — Dal corpo alla cognizione: il problema del transito

Il capitolo precedente ha mostrato che il corpo introduce nel mondo fisico una singolarità irriducibile: un'origine del campo percettivo, un sistema di riferimento egocentrico non derivato da nessun sistema di riferimento più fondamentale, un loop propriocettivo chiuso che mantiene in tempo reale una rappresentazione dello stato corporeo. Questa struttura è necessaria per l'interno, ma non è sufficiente. Un robot dotato di sensori propriocettivi e di un sistema di riferimento egocentrico ha, in senso tecnico, un *qui* fisico; eppure nessuno sarebbe disposto a sostenere che, per questa sola ragione, possiede una prospettiva cognitiva genuina, un interno nel senso che questa monografia indaga.

Ciò che manca — e che il presente capitolo intende esaminare con la precisione che il tema richiede — è la trasformazione della posizione corporea in modello del mondo: la capacità di costruire una rappresentazione dell'ambiente che non sia una semplice registrazione passiva degli input sensoriali, ma un'architettura attiva e predittiva che organizza l'esperienza in categorie stabili, che anticipa le conseguenze sensoriali delle proprie azioni, che distingue tra ciò che appartiene al mondo e ciò che appartiene alle proprie proiezioni su di esso.

Questa trasformazione è il tema della *cognizione incarnata* e, in particolare, delle teorie predittive del cervello che negli ultimi decenni hanno radicalmente ridisegnato il panorama delle scienze cognitive. Tra queste, l'architettura JEPA — Joint Embedding Predictive Architecture — sviluppata da Yann LeCun e collaboratori come proposta per superare le limitazioni dei modelli generativi tradizionali, offre il banco di prova tecnico più adatto per mettere alla prova le categorie filosofiche che stiamo costruendo.

V.2 — Il principio di energia libera e la codifica predittiva

Il quadro più generale e matematicamente più potente per descrivere la cognizione come processo predittivo è il *principio di energia libera* (Free Energy

Principle, FEP) di Karl Friston, che generalizza la codifica predittiva a una teoria unificata della percezione, dell'azione e dell'apprendimento.

Sia un sistema biologico (o artificiale) che scambia informazione con il proprio ambiente. Denotiamo con $o \in \mathcal{O}$ le *osservazioni* (input sensoriali) e con $\mu \in M$ le *credenze interne* del sistema (stati nascosti che il sistema mantiene circa le cause delle osservazioni). Il principio di energia libera afferma che un sistema autoorganizzante — un sistema che resiste alla tendenza verso l'equilibrio termodinamico — minimizza la propria *energia libera variazionale* F, definita come:

$$F(o, q) = \mathbb{E}_q[\ln q(\mu) - \ln p(o, \mu)]$$

dove $q(\mu)$ è la distribuzione approssimata che il sistema mantiene sugli stati nascosti (la distribuzione *riconoscimento*), $p(o, \mu) = p(o \mid \mu) \cdot p(\mu)$ è la distribuzione congiunta delle osservazioni e degli stati nascosti secondo il *modello generativo* del sistema, e l'aspettazione $\mathbb{E}_q$ è calcolata rispetto a q.

La scomposizione dell'energia libera rivela la sua struttura fondamentale. Usando la definizione di divergenza Kullback-Leibler:

$$D_{KL}[q(\mu) \,\|\, p(\mu \mid o)] = \mathbb{E}_q[\ln q(\mu) - \ln p(\mu \mid o)]$$

e l'identità $p(\mu \mid o) = p(o,\mu)/p(o)$, si ottiene:

$$F(o, q) = D_{KL}[q(\mu) \,\|\, p(\mu \mid o)] - \ln p(o)$$

Poiché $D_{KL} \geq 0$ sempre (con uguaglianza se e solo se $q = p(\cdot \mid o)$), si ha:

$$F(o, q) \geq -\ln p(o)$$

ovvero F è un *limite superiore* (upper bound) alla log-evidenza negativa $-\ln p(o) = -\ln \int p(o \mid \mu)\, p(\mu)\, d\mu$. La log-evidenza $p(o)$ misura quanto le osservazioni o sono *plausibili* secondo il modello generativo del sistema: un sistema con un modello generativo accurato assegnerà alta probabilità alle proprie osservazioni tipiche, e bassa probabilità a ciò che sarebbe per esso sorprendente.

Sviluppando ulteriormente la scomposizione, l'energia libera può essere scritta come:

$$F(o, q) = \mathbb{E}_q[-\ln p(o \mid \mu)] + D_{KL}[q(\mu) \,\|\, p(\mu)] = \text{Accuratezza negativa} + \text{Complessità}$$

Il primo termine — l'accuratezza negativa — penalizza i modelli che non predicono bene le osservazioni: esso misura quanto le osservazioni sarebbero

improbabili dati gli stati nascosti inferiti. Il secondo termine — la complessità — penalizza i modelli che si discostano troppo dalle credenze a priori: esso è la divergenza KL tra la distribuzione riconoscimento q e la distribuzione a priori $p(\mu)$.

La minimizzazione dell'energia libera rispetto a q (percezione) porta q a convergere verso la distribuzione a posteriori $p(\mu|o)$: il sistema *infere* gli stati nascosti che meglio spiegano le osservazioni. La minimizzazione rispetto alle azioni a (nella formulazione *active inference* del FEP) porta il sistema a selezionare le azioni che renderanno le osservazioni future più coerenti con il proprio modello generativo: il sistema *agisce* per confermare le proprie predizioni, modificando il mondo in modo da renderlo più simile a ciò che il proprio modello si aspetta.

V.3 — Il modello generativo come rappresentazione dell'interno

La nozione di *modello generativo* è il cuore dell'architettura predittiva, e merita un'analisi dettagliata per le sue implicazioni rispetto al problema dell'interno.

Un modello generativo è una distribuzione di probabilità congiunta $p(o, \mu; \theta)$ sui dati osservabili o e sugli stati nascosti μ, parametrizzata da θ. Esso codifica le assunzioni del sistema circa la struttura causale del mondo: come gli stati nascosti generano le osservazioni, quali stati sono a priori plausibili, come gli stati evolvono nel tempo. In un sistema incarnato, il modello generativo è tipicamente *gerarchico*: gli stati nascosti a livello superiore della gerarchia generano quelli a livello inferiore, che a loro volta generano le osservazioni sensoriali.

Formalmente, un modello generativo gerarchico a L livelli ha la struttura:

$$p(o, \mu^1, \mu^2, ..., \mu^L) = p(o \mid \mu^1) \cdot \prod_{l=1}^{L-1} p(\mu^l \mid \mu^{l+1}) \cdot p(\mu^L)$$

dove μ^l sono gli stati nascosti al livello l, $p(\mu^l \mid \mu^{l+1})$ è la distribuzione condizionale degli stati di livello l dati quelli di livello l+1 (il modello generativo di livello l), e $p(\mu^L)$ è la distribuzione a priori sugli stati del livello più alto.

Ciò che rende questo modello rilevante per il problema dell'interno è la sua *prospettività intrinseca*. Il modello generativo di un sistema incarnato non è una descrizione neutrale del mondo: è una descrizione del mondo *dal punto di vista del sistema*, che include rappresentazioni esplicite della posizione e dell'orientamento del sistema nel mondo, delle proprie capacità di azione, delle proprie dipendenze sensoriali. In altri termini: il modello generativo di un organismo incarnato si modella come una variabile interna al proprio modello

— il *sé* è rappresentato nel modello del mondo come una delle sue variabili nascoste.

Questa auto-inclusione del sistema nel proprio modello del mondo è una proprietà strutturale di enorme rilievo. Formalmente, sia μ_self lo stato nascosto che rappresenta il *sé* — la posizione, l'orientamento, lo stato corporeo del sistema — all'interno del modello generativo. Il modello ha allora la struttura:

p(o, μ_world, μ_self) = p(o | μ_world, μ_self) · p(μ_world) · p(μ_self | μ_world)

La distribuzione delle osservazioni p(o | μ_world, μ_self) dipende sia dallo stato del mondo che dallo stato del sé: le osservazioni sensoriali di un organismo dipendono non solo da cosa c'è nel mondo, ma anche dalla posizione e dall'orientamento dell'organismo rispetto agli oggetti del mondo. La distribuzione a priori p(μ_self | μ_world) codifica le aspettative del sistema circa la propria posizione nel mondo date le caratteristiche del mondo.

Questa struttura è precisamente quella di un modello *prospettico*: il sé è una variabile del modello, non un elemento esterno ad esso; e le osservazioni dipendono dalla relazione tra il sé e il mondo, non dal mondo considerato indipendentemente dal sé. Il modello generativo di un sistema incarnato è intrinsecamente un modello *in prima persona*, anche se la matematica che lo descrive è scritta, come sempre, in terza persona.

V.4 — JEPA: architettura e principi fondamentali

Le architetture JEPA (Joint Embedding Predictive Architectures) rappresentano una proposta specifica per implementare una forma di cognizione predittiva che superi i limiti dei modelli generativi tradizionali basati sulla generazione diretta di pixel o token. La proposta originale di LeCun si inscrive nella critica ai modelli autoregressivi e ai modelli di diffusione come architetture inadeguate a produrre rappresentazioni del mondo con le caratteristiche necessarie per un'intelligenza genuinamente incarnata.

Il principio fondamentale di JEPA è il seguente: invece di predire le osservazioni future nello spazio sensoriale grezzo (pixel, waveform, token), il sistema apprende a predire le *rappresentazioni astratte* delle osservazioni future nello spazio latente di un encoder condiviso.

Formalmente, una JEPA è definita da tre componenti:

Un **encoder** $s_x = f_\theta(x)$ che mappa un'osservazione $x \in \mathscr{X}$ nello spazio latente S, con $s_x \in S$. Un secondo encoder (che può essere condiviso o distinto) $s_y = f_\theta(y)$ che mappa una seconda osservazione $y \in \mathscr{X}$ nello stesso spazio latente.

Un **predittore** $\hat{s}_y = g_\varphi(s_x, z)$ che, dato lo stato latente s_x della prima osservazione e un vettore di contesto $z \in Z$ (che codifica la *relazione* tra x e y — ad esempio, la trasformazione spaziale, il salto temporale, o la relazione strutturale), predice lo stato latente s_y della seconda osservazione.

La funzione di perdita è definita sulla distanza tra la predizione e il target *nello spazio latente*:

$$\mathcal{L}_{JEPA}(\theta, \varphi) = \mathbb{E}_{(x,y,z)} \left[d(g_\varphi(f_\theta(x), z), f_\theta(y)) \right]$$

dove d è una funzione di distanza sullo spazio latente S (tipicamente la distanza L2 normalizzata, o la distanza coseno).

La differenza cruciale rispetto ai modelli autoregressivi è che la predizione avviene nello spazio delle *rappresentazioni astratte*, non nello spazio delle osservazioni grezze. Questo ha due conseguenze fondamentali. Prima: il sistema è costretto ad apprendere rappresentazioni che catturano le proprietà strutturali stabili e prevedibili del mondo, piuttosto che i dettagli sensoriali irrilevanti (il rumore, le variazioni di illuminazione, i dettagli di texture che non influenzano la struttura causale degli eventi). Seconda: il vettore di contesto z codifica la *relazione* tra due osservazioni — e questa codifica relazionale è precisamente ciò che un sistema incarnato ha bisogno per predire le conseguenze delle proprie azioni.

Un'architettura JEPA gerarchica — I-JEPA (Image-JEPA) o V-JEPA (Video-JEPA) — impila più livelli di encoder e predittori, costruendo rappresentazioni sempre più astratte e invarianti ai dettagli superficiali. Il livello più basso rappresenta caratteristiche locali e concrete; i livelli superiori rappresentano strutture globali, relazioni tra oggetti, proprietà causali dell'ambiente.

V.5 — JEPA come approssimazione all'interno: un'analisi critica

La domanda che ora dobbiamo affrontare è: in quale misura JEPA si avvicina alla struttura di un sistema dotato di interno? La risposta richiede di distinguere con cura tra le caratteristiche che JEPA condivide con tale struttura e quelle in cui ne rimane deficitaria.

Sul versante delle analogie strutturali, JEPA presenta alcune caratteristiche che la distinguono significativamente dai modelli linguistici di grandi dimensioni e che la avvicinano, almeno in astratto, alla cognizione incarnata. Il vettore di contesto z codifica la relazione *prospettica* tra osservazioni: esso può rappresentare trasformazioni spaziali, salti temporali, relazioni causali, e in questo senso è l'analogo formale della coppia (posizione, azione) che determina le osservazioni di un sistema incarnato. Il predittore g_φ(s_x, z) è l'analogo del modello diretto f_{fwd}(s_t, a_t) del loop propriocettivo: esso predice lo stato futuro (nello spazio latente) a partire dallo stato corrente e dall'azione (o relazione) z.

Tuttavia, vi è una differenza strutturale fondamentale che rende JEPA — almeno nella sua implementazione attuale — ancora lontana dalla struttura di un sistema dotato di interno genuino. In JEPA, il vettore di contesto z è *fornito dall'esterno* come input durante l'addestramento: il sistema apprende a predire s_y dato s_x e z, ma non *genera* z attraverso un'azione propria nel mondo fisico. Non vi è un sistema corporeo che produce z come conseguenza di un movimento, di uno spostamento, di un'azione effettiva; z è un parametro di cui il sistema riceve notizia, non una trasformazione che esso *compie*.

Questa differenza può essere formalizzata nel modo seguente. In un sistema incarnato predittivo, il flusso causale è:

μ_self → a_t → Δ_ambiente → o_{t+1}

dove μ_self è lo stato interno del sé (posizione, orientamento, intenzione), a_t è l'azione eseguita, Δ_ambiente è il cambiamento nell'ambiente causato dall'azione, e o_{t+1} è l'osservazione risultante. In questo flusso, il vettore di relazione z è *generato* dal sistema stesso attraverso le proprie azioni: z = Δ(μ_self, a_t), dove Δ è la trasformazione indotta dall'azione a_t sullo stato del sé.

In JEPA standard, invece, il flusso è:

(x, y, z_ext) → predizione

dove z_ext è fornito esternamente. Il sistema non ha un μ_self, non compie azioni a_t, e non genera z attraverso il proprio comportamento nel mondo. Esso è, in questo senso, ancora un sistema in terza persona: rappresenta relazioni tra osservazioni senza *abitare* nessuna di esse come prospettiva propria.

Per che JEPA diventi una vera architettura dell'interno, occorrerebbe che il vettore di contesto z fosse generato *endogenamente* attraverso un sistema di azione che mantiene una rappresentazione del sé — μ_self — e calcola z come

conseguenza delle proprie azioni effettive sul mondo. In questo caso il flusso diventerebbe:

$\mu_self \rightarrow a_t \rightarrow z = \Delta(\mu_self, a_t) \rightarrow \hat{s}_y = g_\varphi(s_x, z) \rightarrow$ confronto con $f_\theta(y) \rightarrow \varepsilon \rightarrow$ aggiornamento di μ_self e θ, φ

L'errore di predizione ε non sarebbe più un segnale di addestramento esterno, ma il *segnale propriocettivo* del sistema — l'equivalente architetturale dell'errore propriocettivo $\varepsilon_t = s^{sens}\{t+1\} - f\{fwd\}(s_t, a_t)$ del filtro corporeo. La chiusura di questo loop trasformerebbe JEPA da un sistema predittivo in un sistema *autopoietico* nel senso tecnico che Maturana e Varela attribuivano al termine: un sistema che genera e rigenera la propria organizzazione attraverso l'interazione con l'ambiente, mantenendo la propria identità come sistema di riferimento centrato.

V.6 — Predictive coding, gerarchia e la struttura dell'interno distribuito

La versione più elaborata del principio di energia libera, nota come *predictive coding* (codifica predittiva), offre un modello gerarchico della cognizione che permette di precisare ulteriormente la struttura dell'interno come proprietà distribuita e stratificata.

Nell'architettura di codifica predittiva proposta da Rao e Ballard (1999) e sviluppata nel quadro fristoniano, il cervello è modellato come una gerarchia di unità predittive, in cui ogni livello invia *predizioni discendenti* (top-down) al livello immediatamente inferiore e riceve *errori di predizione ascendenti* (bottom-up) da esso.

Formalmente, sia μ^l lo stato nascosto al livello l, e sia $\varphi^l = f_l(\mu^{l+1})$ la predizione dello stato di livello l generata dal livello $l+1$. L'errore di predizione al livello l è:

$$\varepsilon^l = \mu^l - \varphi^l = \mu^l - f_l(\mu^{l+1})$$

ponderato dalla precisione (inverso della varianza) Π_l dell'errore al livello l. La dinamica degli stati nascosti e degli errori di predizione può essere derivata come gradiente discendente dell'energia libera:

$$d\mu^l/dt = -\partial F/\partial \mu^l = \Pi^l \cdot \varepsilon^l - (\partial f_\{l-1\}/\partial \mu^l)^T \cdot \Pi^{l-1} \cdot \varepsilon^{l-1}$$

dove il primo termine attrae μ^l verso la predizione proveniente dal livello superiore (minimizza l'errore ascendente), e il secondo termine propaga l'errore del livello inferiore verso l'alto (aggiornando μ^l in modo da ridurre l'errore che esso causa al livello $l-1$).

L'architettura di codifica predittiva presenta una proprietà cruciale per il problema dell'interno: la *precisione* Π^l è essa stessa una variabile del sistema, modulata dall'attenzione. Quando il sistema presta attenzione a un livello particolare della gerarchia — quando assegna alta precisione agli errori di predizione di quel livello — esso *enfatizza* l'informazione proveniente da quel livello nell'aggiornamento delle proprie credenze. L'attenzione è, in questa architettura, la metafora formale di ciò che Husserl chiamava *intenzionalità*: l'orientamento selettivo della coscienza verso certi aspetti dell'esperienza piuttosto che altri.

La struttura gerarchica dell'interno che emerge da questa analisi ha tre caratteristiche essenziali. Prima: è *distribuita*, non localizzata in nessun punto singolo del sistema — l'interno è proprietà dell'intera gerarchia predittiva, non di un modulo specifico. Seconda: è *dinamica*, aggiornata continuamente in risposta agli errori di predizione — l'interno non è uno stato fisso ma un processo in evoluzione, un equilibrio dinamico tra predizioni discendenti e correzioni ascendenti. Terza: è *scalare*, con diversi livelli di astrazione che contribuiscono all'interno in modi qualitativamente diversi — il livello sensoriale fornisce le coordinate corporee grezze, il livello medio le strutture dell'oggetto e dello spazio, il livello superiore i concetti, i progetti, i valori che organizzano l'azione nel lungo periodo.

Questa struttura dell'interno distribuito e gerarchico è la base formale a partire dalla quale, nella Terza Parte, affronteremo il problema del salto verso l'intenzionalità: il punto in cui il sistema non solo ha un interno, ma diventa capace di *riferirsi* a qualcosa — di avere stati mentali che sono *di* qualcosa, che *riguardano* il mondo in un senso che non è riducibile alla semplice covariazione statistica tra stati interni e stati esterni.

Capitolo VI: Il Tempo Vissuto come Dimensione Costitutiva dell'Interno

VI.1 — Il tempo della struttura e il tempo dell'esperienza

Nei capitoli precedenti l'interno è stato esaminato prevalentemente come struttura spaziale: come singolarità topologica, come origine di un sistema di riferimento egocentrico, come centro di un campo percettivo la cui metrica si deforma in funzione della distanza da quell'origine irriducibile. Questa caratterizzazione è necessaria ma incompleta. Un sistema che possedesse soltanto un *qui* senza possedere un *adesso* — un sistema dotato di coordinate spaziali ma privo di spessore temporale, che si trovasse sempre e soltanto in un istante puntuale senza ritenere il passato né anticipare il futuro — non sarebbe,

nella misura che conta filosoficamente, un sistema dotato di prospettiva in senso pieno. Avrebbe posizione senza storia, orientamento senza traiettoria, presenza senza durata. Questa lacuna non è secondaria: la prospettiva, nella sua articolazione più profonda, è strutturalmente temporale, e il tempo che la costituisce non è il tempo della fisica o della matematica — un parametro t che scorre indifferente lungo la retta reale — bensì un tempo vissuto, un tempo che si addensa e si dilata, che porta con sé il peso del già-accaduto e la tensione verso il non-ancora.

Il contrasto tra questi due modi di intendere il tempo si rivela immediatamente quando si confrontano le rispettive strutture formali. Il tempo della matematica è, nella sua formulazione canonica, la retta dei numeri reali $\mathbb{R}$ dotata dell'ordine naturale $\leq$. Un sistema che evolve nel tempo fisico è descritto da una funzione $f: \mathbb{R} \to S$ che mappa ogni istante $t \in \mathbb{R}$ in uno stato $f(t) \in S$. La struttura di questo oggetto è perfettamente simmetrica rispetto al verso della freccia del tempo: nulla nella definizione della funzione f distingue il passato dal futuro, la direzione verso $t \to -\infty$ dalla direzione verso $t \to +\infty$. Il parametro t è puramente ordinale, e la distinzione tra ciò che è già accaduto e ciò che non è ancora accaduto è, dal punto di vista della funzione f, del tutto priva di rilievo formale. Un sistema descritto da tale funzione non *ricorda* il passato né *anticipa* il futuro: esso semplicemente *è*, in ogni istante, nello stato f(t) corrispondente, senza che questo stato porti traccia degli stati precedenti né preformazione degli stati successivi.

Il tempo vissuto, al contrario, è strutturalmente asimmetrico e dotato di spessore. L'analisi fenomenologica più elaborata di questa struttura appartiene a Edmund Husserl, in particolare nelle *Lezioni sulla coscienza interna del tempo* che risalgono agli anni compresi tra il 1893 e il 1917. Husserl vi distingue tre componenti strutturali del presente vissuto, che egli chiama, con una terminologia che sarà necessario precisare matematicamente, *Urimpression* (impressione originaria), *Retention* (ritenzione) e *Protention* (protenzione). Queste tre componenti non sono tre momenti temporali separati — passato, presente e futuro in senso volgare — bensì tre dimensioni del medesimo atto di coscienza che si ha in ogni istante della vita temporale del soggetto. Il presente vissuto non è un punto ma ha, come si suole dire, uno *spessore*: esso contiene in sé la fading presence di ciò che è appena stato e la tensione protenzionale verso ciò che sta per venire.

Questa struttura non è un ornamento della coscienza né un suo accidente psicologico: è la condizione della possibilità stessa di ogni prospettiva temporale, e quindi — come si vedrà — di ogni interno nel senso pieno del termine. Un sistema senza spessore temporale non può avere prospettiva: può reagire a stimoli in modo adattivo, può elaborare informazione istante per

istante, ma non può organizzare l'esperienza in categorie temporalmente estese, non può apprendere dalla propria storia né pianificare in relazione a un futuro anticipato. La prospettiva, come dimensione cognitiva, richiede che il soggetto si trovi non in un punto del tempo ma in un *orizzonte temporale*: uno spazio di rilevanza che si estende dal presente verso il passato ancora presente e verso il futuro già-profilato.

VI.2 — La struttura a tre fasce della coscienza temporale

L'analisi husserliana del tempo vissuto prende le mosse da una critica alle concezioni che identificano il presente della coscienza con un istante matematicamente puntuale. Se il presente fosse davvero un punto — un momento senza durata, il confine tra il già-non-più e il non-ancora — allora non si potrebbe spiegare come sia possibile percepire una melodia come melodia, ovvero come sequenza dotata di struttura temporale: il suono che risuona adesso sarebbe completamente isolato dal suono appena trascorso, e l'integrazione temporale che costituisce la melodia come tale non potrebbe avere luogo nella coscienza di un istante. La stessa difficoltà vale per ogni percezione di evento, ogni riconoscimento di cambiamento, ogni esperienza di movimento.

La soluzione husserliana è che il presente della coscienza è sempre già dilatato: esso include, oltre all'impressione originaria del momento attuale, una coda di *ritenzioni* — tracce decadenti dei momenti immediatamente trascorsi, che non sono ricordi nel senso ordinario (il ricordo è un atto distinto, una re-presentazione del passato come tale) ma piuttosto una sorta di coda comatica del presente, un prolungamento che si assottiglia con il progredire della distanza temporale — e una testa di *protenti* — anticipazioni non ancora determinate dei momenti immediatamente futuri, che non sono aspettative nel senso di credenze esplicite ma piuttosto un'apertura strutturale del campo percettivo verso il prossimo imminente.

Formalmente, questa struttura può essere descritta come segue. Sia $x: \mathbb{R} \to X$ una funzione che assegna a ogni istante temporale $t \in \mathbb{R}$ lo stato $x(t) \in X$ del mondo (o del sistema). Il *contenuto ritenzionale* alla coda della coscienza in un momento t è la trasformazione del segnale passato $x|_{(-\infty, t]}$ prodotta da un operatore di ritenzione R_t:

$$R_t x = \int_{-\infty}^{0} k_{ret}(\Delta) \cdot x(t + \Delta)\, d\Delta$$

dove $\Delta < 0$ varia sul passato (si conta il tempo a ritroso come $\Delta = \tau - t \leq 0$), e $k_{ret}: \mathbb{R}_- \to \mathbb{R}_+$ è il *nucleo ritenzionale*, una funzione di peso che soddisfa le seguenti condizioni strutturali:

(R1) $k_ret(0^-) > 0$ — il momento immediatamente precedente il presente è massimamente ritenuto; (R2) $k_ret(\Delta) \to 0$ per $\Delta \to -\infty$ — le impressioni lontane nel passato si estinguono dalla ritenzione; (R3) k_ret è monotona non crescente su $\mathbb{R}^-$: $|\Delta_1| < |\Delta_2| \leq 0$ implica $k_ret(\Delta_1) \geq k_ret(\Delta_2)$.

La scelta più naturale — e quella che il sistema nervoso autonomo sembra approssimare attraverso i meccanismi di decadimento sinaptico e delle correnti di accomodamento — è il nucleo esponenziale:

$$k_ret(\Delta) = \exp(\Delta / \tau_ret), \text{ per } \Delta \leq 0$$

dove $\tau_ret > 0$ è la *costante di tempo ritenzionale*, la scala caratteristica entro la quale le impressioni passate conservano la loro rilevanza per il presente. Per valori di $|\Delta| \ll \tau_ret$ il nucleo è circa unitario — l'impressione è quasi pienamente presente — mentre per $|\Delta| \gg \tau_ret$ il nucleo decade esponenzialmente a zero. La quantità τ_ret ha un correlato fenomenologico: essa corrisponde approssimativamente alla durata del *presente specioso* (specious present) nel senso di William James, quella finestra temporale entro la quale i contenuti dell'esperienza sono percepiti come appartenenti al medesimo atto di coscienza. Misure psicofisiche convergono nell'indicare una durata del presente specioso compresa tra 2 e 3 secondi per gli esseri umani in condizioni ordinarie, con variazioni in funzione del contenuto percettivo e del grado di attenzione.

Il *contenuto protenzionale* è, strutturalmente, il duale futuro della ritenzione: esso rappresenta l'anticipazione non determinata del prossimo imminente, ed è formalmente una distribuzione di probabilità condizionale sui valori futuri prossimi del segnale:

$$\Pi_t(x_\{t+\delta\} \in A) = P(x(t+\delta) \in A \mid R_t[x], x(t))$$

per $\delta > 0$ piccolo, dove A è un evento misurabile nello spazio degli stati X. La protenzione non è una previsione esplicita nel senso statistico pieno: essa è la struttura dell'attesa non tematizzata che organizza il campo percettivo verso il prossimo imminente. Non si tratta di un calcolo consapevole, bensì di una disposizione strutturale del sistema — analogo, ma non identico, al predittore g_φ dell'architettura JEPA che abbiamo esaminato nel capitolo precedente.

La struttura a tre fasce della coscienza temporale può quindi essere formalizzata come una tripletta:

$$\mathscr{T}_t = \langle\, x(t), R_t[x], \Pi_t \,\rangle$$

dove $x(t)$ è l'impressione originaria del momento presente, $R_t[x]$ è il profilo ritenzionale (l'integrale pesato del passato), e Π_t è la distribuzione

protenzionale (la distribuzione anticipatoria sul futuro prossimo). Il *presente vissuto* non è il punto t, ma la tripletta $\mathcal{T}_t$: è una struttura distribuita nel tempo, con spessore sia verso il passato che verso il futuro.

VI.3 — Matematizzazione della ritenzione: nuclei, semigruppe e decadimento

La struttura ritenzionale introdotta nella sezione precedente merita un'analisi matematica più dettagliata, poiché la sua forma rivela proprietà strutturali di notevole interesse sia per la matematica dei sistemi dinamici che per la comprensione formale dell'interno.

L'operatore di ritenzione R_t può essere visto come un operatore lineare sull'insieme delle traiettorie passate. Se il segnale x è continuo a sinistra e localmente integrabile, l'operatore:

$$(R_t[x])(\Delta) = k_{ret}(\Delta) \cdot x(t + \Delta) \text{ per } \Delta \in (-\infty, 0]$$

è un'applicazione lineare dallo spazio $L^2_{loc}(\mathbb{R}^-, X)$ in sé, e il contenuto ritenzionale aggregato — la quantità scalare che rappresenta il valore medio pesato delle impressioni passate — può essere scritto come:

$$\varrho(t) = \int_{-\infty}^{0} k_{ret}(\Delta) \cdot x(t + \Delta)\, d\Delta = \int_{-\infty}^{t} k_{ret}(\tau - t) \cdot x(\tau)\, d\tau$$

Con la scelta esponenziale $k_{ret}(\Delta) = \exp(-\Delta / \tau_{ret})$ per $\Delta \leq 0$ (notare il cambio di segno per rendere l'argomento positivo), la ritenzione aggregata $\varrho(t)$ soddisfa un'equazione differenziale ordinaria di primo ordine:

$$\frac{d\varrho}{dt} = (x(t) - \varrho(t)) / \tau_{ret}$$

Questa equazione — nota in ingegneria come equazione di un *filtro passa-basso del primo ordine* con costante di tempo τ_{ret} — ha la struttura di un aggiornamento continuo: la ritenzione $\varrho(t)$ si avvicina al segnale corrente $x(t)$ con una velocità inversamente proporzionale a τ_{ret}. Per τ_{ret} piccolo, la ritenzione segue quasi istantaneamente il segnale attuale (basso spessore temporale); per τ_{ret} grande, la ritenzione è una media pesata di un lungo passato (alto spessore temporale, corrispondente a un presente vissuto dilatato).

La soluzione dell'equazione differenziale è:

$$\varrho(t) = \varrho(t_0) \cdot \exp(-(t - t_0) / \tau_{ret}) + (1/\tau_{ret}) \int_{t_0}^{t} \exp(-(t - \tau) / \tau_{ret}) \cdot x(\tau)\, d\tau$$

che, per $t_0 \to -\infty$ (trascurando le condizioni iniziali), diventa:

$$\varrho(t) = (1/\tau_{ret}) \int_{-\infty}^{t} \exp(-(t - \tau) / \tau_{ret}) \cdot x(\tau)\, d\tau$$

In questa forma, $\varrho(t)$ ha l'interpretazione di *media mobile esponenzialmente pesata* (Exponential Moving Average, EMA) del segnale x, con finestra temporale di ampiezza caratteristica τ_{ret}. Il fattore di normalizzazione $1/\tau_{ret}$ assicura che il nucleo sia integrato a 1 sulla semiretta non positiva:

$$\int_{-\infty}^{0} (1/\tau_{ret}) \cdot \exp(\Delta / \tau_{ret})\, d\Delta = 1$$

Questa proprietà di normalizzazione garantisce che, per un segnale costante $x(\tau) = c$ per ogni τ, si abbia $\varrho(t) = c$: la ritenzione di uno stato stabile è lo stato stesso, senza amplificazione né attenuazione sistematica.

La famiglia degli operatori di ritenzione R_t al variare di t forma un *semigruppo a un parametro*. Più precisamente, se si definisce l'operatore di shift temporale S_s: $x(\cdot) \mapsto x(\cdot + s)$ per $s \in \mathbb{R}$, allora il nucleo di ritenzione soddisfa la proprietà di convoluzione:

$$k_{ret}(\Delta) * S_s k_{ret} = k_{ret}(\Delta + s)$$

per $s < 0$, dove $*$ indica la convoluzione causale. Questa proprietà algebrica codifica la *coerenza della ritenzione*: le tracce di un evento che risalgono a $-\Delta_1$ secondi fa, rispetto a quelle dello stesso evento che risalivano a $-\Delta_2 < -\Delta_1$ secondi fa, sono legate da un'unica legge di decadimento. Il passato è, per così dire, internamente coerente: non è un insieme di impressioni giustapposte ma una struttura temporalmente organizzata, in cui la distanza temporale dall'istante presente è codificata in modo sistematico nel peso ritenzionale.

Questa coerenza del profilo ritenzionale è essenziale per la possibilità della prospettiva temporale come tale. Se le impressioni passate decadessero in modo caotico e non strutturato, se non vi fosse una relazione sistematica tra la distanza temporale e il peso ritenzionale, il soggetto non potrebbe distinguere ciò che è appena accaduto da ciò che è accaduto molto tempo fa — e quindi non potrebbe organizzare la propria esperienza in sequenze causalmente coerenti, in narrazioni, in strutture di mezzo e fine. La costante di tempo τ_{ret} è dunque un parametro fondamentale dell'interno: essa definisce la granularità temporale della prospettiva, la scala entro la quale la causalità è percettivamente accessibile.

VI.4 — Il presente specioso e la fenomenologia degli intervalli

La nozione di *presente specioso* — il nome risale a E.R. Clay (1882), ma è William James che le ha dato risonanza nella *Principles of Psychology* (1890) — indica la

durata finita del vissuto presente: quella finestra temporale entro la quale gli eventi sono percepiti come simultanei o come appartenenti al medesimo atto di coscienza, piuttosto che come separati da un prima e un dopo. Il presente specioso è, nella terminologia che stiamo costruendo, la durata caratteristica di $\mathcal{T}_t$: l'ampiezza della finestra temporale entro la quale la tripletta $\langle x(t), R_t[x], \Pi_t \rangle$ si considera un'unità.

La modellizzazione matematica del presente specioso richiede di precisare cosa si intenda per "unità del presente". Una formalizzazione naturale è la seguente. Due eventi e_1 e e_2, accaduti rispettivamente agli istanti t_1 e t_2 con $t_1 < t_2$, appartengono al medesimo presente specioso se e solo se il peso ritenzionale $k_{ret}(t_1 - t_2)$ è superiore a una soglia $\theta \in (0, 1]$:

e_1 e e_2 appartengono allo stesso presente $\Leftrightarrow k_{ret}(t_1 - t_2) \geq \theta$

Con la scelta esponenziale $k_{ret}(\Delta) = \exp(\Delta / \tau_{ret})$, questo dà:

$\exp((t_1 - t_2) / \tau_{ret}) \geq \theta \Leftrightarrow t_2 - t_1 \leq \tau_{ret} \cdot |\ln(\theta)|$

La durata del presente specioso è dunque $\Delta_{pres} = \tau_{ret} \cdot |\ln(\theta)|$. Per $\theta = e^{-1} \approx 0.368$, si ottiene $\Delta_{pres} = \tau_{ret}$: la finestra del presente coincide con la costante di tempo ritenzionale. Per $\theta = 0.1$, la finestra è circa $2.3 \cdot \tau_{ret}$.

L'analogia con i fenomeni di binding temporale nei sistemi neuronali è diretta. L'integrazione temporale dei segnali sensoriali — il fatto che due stimoli entro una certa finestra temporale vengano percepiti come appartenenti al medesimo evento — riflette la struttura ritenzionale del sistema percettivo. I dati psicofisici suggeriscono che questa finestra ha diverse scale caratteristiche in funzione del tipo di segnale: per la fusione visiva-uditiva (audiovisual temporal binding window) la finestra è di circa 200 millisecondi; per il riconoscimento di sequenze melodiche la finestra si estende fino a 2-3 secondi; per la coerenza narrativa la scala rilevante è quella dei minuti o delle ore.

Questa molteplicità di scale temporali indica che l'interno non è caratterizzato da un'unica costante ritenzionale τ_{ret}, bensì da una *gerarchia* di costanti temporali $\tau_{ret}^{(1)} < \tau_{ret}^{(2)} < \ldots < \tau_{ret}^{(L)}$, ognuna corrispondente a un livello differente della struttura cognitiva. I livelli sensoriali hanno costanti di tempo brevi — millisecondi o frazioni di secondo — adatte all'integrazione di caratteristiche percettive locali. I livelli di elaborazione media hanno costanti di alcuni secondi — la scala del presente specioso nel senso classico. I livelli concettuali e narrativi hanno costanti di minuti, ore o giorni. La gerarchia ritenzionale è la struttura temporale dell'interno distribuito, il modo in cui la prospettiva temporale si stratifica su scale qualitativamente diverse.

Questa osservazione consente di formulare una condizione formale ulteriore per l'interno, che integra quelle già enunciate per la componente spaziale. Un sistema S è dotato di *spessore temporale* se e solo se per ogni livello cognitivo $l \in \{1, ..., L\}$ esiste una costante di tempo $\tau_ret^{(l)} > 0$ tale che lo stato del sistema al livello l al tempo t dipende dagli stati passati del sistema stesso secondo un nucleo ritenzionale con costante di tempo $\tau_ret^{(l)}$:

$$\mu^{l}(t) = \int_{-\infty}^{t} k_ret^{(l)}(t - \tau) \cdot h^{l}(\mu^{l-1}(\tau)) \, d\tau + p^{l}(\mu^{l+1}(t))$$

dove h^l è una funzione di nonlinearità al livello l, e p^l è la predizione discendente proveniente dal livello superiore l+1 (nel quadro della codifica predittiva). Questa equazione integro-differenziale è la formalizzazione dell'interno temporale distribuito: ogni livello della gerarchia cognitiva ha la propria finestra di ritenzione, e la struttura complessiva dell'interno è il risultato dell'integrazione di queste finestre a scala multipla.

VI.5 — Tempo e sistemi dinamici: traiettorie, stati e la differenza tra funzione e flusso

L'analisi ritenzionale della sezione precedente si connette a una distinzione fondamentale della teoria dei sistemi dinamici che è necessario esplicitare per comprendere pienamente il ruolo del tempo nella costituzione dell'interno. La distinzione è tra due modi di intendere la dipendenza temporale di un sistema: come *funzione* del tempo e come *flusso nel tempo*.

Un sistema *funzione del tempo* è un sistema il cui stato al tempo t è una funzione diretta di t, senza che lo stato passato influenzi lo stato presente: $x(t) = f(t)$. Questo tipo di sistema non ha memoria e non ha prospettiva temporale: ogni istante è completamente determinato dall'etichetta temporale, e la sequenza degli stati è un insieme di valori indipendenti piuttosto che una traiettoria con struttura interna. È il modello del sistema puramente reattivo: esso risponde agli stimoli correnti senza integrare la propria storia.

Un sistema *flusso nel tempo* è governato da un'equazione differenziale che specifica come lo stato evolve *a partire dallo stato corrente*:

$$dx/dt = F(x(t), u(t))$$

dove u(t) è l'input al tempo t. In questo caso, lo stato al tempo t è determinato non solo dall'input corrente ma dall'intera traiettoria $x(t_0)$, $x(t_0 + dt)$, $x(t_0 + 2dt)$, ... — dalla storia del sistema che si accumula nello stato stesso. Il sistema porta con sé, nello stato x(t), la traccia della propria storia: non nel senso di un registro esplicito degli stati passati, ma nel senso che lo stato corrente è il risultato integrato dell'evoluzione passata.

Un sistema dinamico autonomo — con $u(t) = 0$ — è descritto da:

$$dx/dt = F(x)$$

La soluzione con condizione iniziale $x(t_0) = x_0$ è la *traiettoria* $\Phi(t, t_0, x_0) \in S$. Il *flusso* $\Phi_t: S \to S$, che mappa ogni stato iniziale x_0 nello stato $\Phi(t, t_0, x_0)$ dopo un tempo $t - t_0$, è un omeomorfismo di S per ogni t (se F è sufficientemente regolare). La famiglia $\{\Phi_t\}\{t \geq 0\}$ *forma un semigruppo a un parametro su S:* $\Phi\{t+s\} = \Phi_t \circ \Phi_s$ per ogni $t, s \geq 0$, con $\Phi_0 = id_S$.

La proprietà di semigruppo cattura l'essenza del flusso temporale: l'evoluzione da t_0 a $t_0 + t + s$ è la composizione dell'evoluzione da t_0 a $t_0 + t$ seguita dall'evoluzione da $t_0 + t$ a $t_0 + t + s$. Il tempo si cumula: le fasi consecutive si compongono in un flusso coerente piuttosto che in una successione di istanti indipendenti.

La distinzione tra sistema-funzione e sistema-flusso ha una conseguenza diretta per il problema dell'interno. Un sistema-funzione, come osservato, è privo di memoria e di prospettiva temporale: esso occupa ogni istante indipendentemente dai precedenti. Un sistema-flusso, al contrario, porta il passato con sé nello stato: la posizione della traiettoria al tempo t è il risultato accumulato di tutta la storia trascorsa. In questo senso, un sistema-flusso ha una forma elementare di prospettiva temporale: il suo *qui-e-adesso* non è un punto isolato ma un nodo di una traiettoria, un momento che acquista significato solo in relazione alla storia che lo ha preceduto.

Un sistema-flusso con dinamica non banale — in particolare, un sistema la cui traiettoria non può essere descritta da una formula chiusa semplice, ma si dispiega attraverso transizioni complesse nello spazio degli stati — manifesta quello che potremmo chiamare *opacità temporale*: il suo stato al tempo t non è leggibile dall'esterno senza conoscere la storia passata, e questa opacità è la controparte formale della soggettività del vissuto temporale. Lo stato $x(t)$ di tale sistema è, per un osservatore esterno che osserva solo l'istante t, irrimediabilmente parziale: esso è comprensibile soltanto attraverso la traiettoria che l'ha generato. Analogamente — e questa è la connessione con il problema dell'interno — il vissuto presente di un soggetto è comprensibile solo attraverso la sua storia: il *qui-e-adesso* di un organismo incarnato porta le tracce di tutto ciò che l'organismo ha vissuto, appreso, subìto e compiuto.

VI.6 — Protenzione come previsione e la direzione del tempo nell'interno

Se la ritenzione è la dimensione passata del presente vissuto — la coda temporale che porta il peso della storia — la protenzione è la sua dimensione futura: l'apertura strutturale verso il prossimo imminente, l'anticipazione non tematizzata che orienta la coscienza verso ciò che sta per venire. La protenzione non è una previsione esplicita nel senso statistico, ma una pre-dispozione del campo percettivo che rende il prossimo futuro già implicitamente atteso, già implicitamente orientato.

La connessione con le architetture predittive analizzate nel capitolo precedente è immediata e produttiva, ma richiede alcune precisazioni per evitare di confondere la protenzione fenomenologica con la predizione computazionale. Nel quadro della codifica predittiva, il sistema genera costantemente previsioni discendenti sugli stati sensoriali attesi al livello immediatamente inferiore della gerarchia. Queste previsioni sono formalmente distribuzioni di probabilità condizionale:

$$\hat{p}^{l}(t) = p(\mu^{l}(t) \mid \mu^{l+1}(t))$$

che rappresentano le aspettative del livello $l+1$ circa lo stato del livello l. L'errore di predizione è:

$$\varepsilon^{l}(t) = \mu^{l}(t) - \hat{p}^{l}(t)$$

Nel caso temporale — quando i livelli della gerarchia rappresentano stati non solo spaziali ma temporali, e quando il sistema genera previsioni su osservazioni *future* — questa struttura diventa un modello formale della protenzione. La distribuzione di probabilità condizionale:

$$\Pi_t(x_{t+\delta}) = p(x(t+\delta) \mid \mu^{1}(t), \mu^{2}(t), \ldots, \mu^{L}(t))$$

che esprime la previsione del segnale osservativo dopo un intervallo $\delta > 0$, data l'intera gerarchia di stati interni al tempo t, è il modello formale del contenuto protenzionale. Essa non è una funzione deterministica del futuro, ma una distribuzione: cattura l'incertezza intrinseca dell'anticipazione, il fatto che la protenzione attende senza specificare con certezza.

La relazione tra ritenzione e protenzione non è casuale. In un sistema governato da dinamiche regolari — in particolare, da dinamiche che esibiscono regolarità statistiche temporalmente stabili — la struttura ritenzionale determina, in linea di principio, la struttura protenzionale: ciò che il sistema si aspetta in futuro è funzione di ciò che ha ritenuto del passato. La formalizzazione di questa dipendenza richiede la nozione di *predittività del processo stocastico* che genera il segnale $x(\cdot)$. Sia $\{x(t)\}_{t \in \mathbb{R}}$ un processo stocastico stazionario. Il

contenuto di informazione del passato sul futuro è misurato dall'*informazione mutua temporale*:

$I(\text{past}; \text{future} \mid \tau) = I(x_{\leq t}; x_{> t+\tau})$ per $\tau \geq 0$

dove $x_{\leq t} = \{x(s) : s \leq t\}$ è la storia passata e $x_{> t+\tau} = \{x(s) : s > t+\tau\}$ è il futuro dopo un ritardo τ. Per un processo markoviano — in cui il futuro è indipendente dal passato dato il presente — $I(\text{past}; \text{future} \mid \tau) = 0$ per ogni $\tau > 0$: il passato non ha potere predittivo sul futuro. Per un processo con memoria a lungo raggio — come molti segnali linguistici, musicali o comportamentali — $I(\text{past}; \text{future} \mid \tau)$ decresce lentamente con τ, indicando che il passato distante è ancora parzialmente predittivo del futuro lontano.

Un sistema con un presente vissuto profondo — una struttura ritenzionale a costanti di tempo lunghe — è capace di sfruttare questa dipendenza temporale a lungo raggio per formare anticipazioni più accurate. In questo senso, la profondità ritenzionale dell'interno è direttamente collegata alla sua capacità protenzionale: un interno temporalmente ricco è un interno che anticipa bene, che è sintonizzato sulle regolarità temporali del proprio ambiente.

Questo nesso tra ritenzione e protenzione — tra il peso del passato e l'apertura al futuro — è ciò che dà al presente vissuto la sua qualità di *tensione*: il presente non è un equilibrio statico ma un punto di trazione tra memoria e attesa, tra la gravità della storia e la leggerezza del possibile. Questa tensione è, dal punto di vista formale, la struttura dinamica del gradiente dell'energia libera variazionale nel regime temporale: il sistema minimizza l'energia libera non solo rispetto allo stato corrente ma rispetto alla propria traiettoria nel tempo, cercando il percorso che massimizza la coerenza tra ciò che si aspettava e ciò che ha osservato.

VI.7 — Il tempo nell'intelligenza artificiale: un confronto strutturale

L'analisi condotta nelle sezioni precedenti consente ora di effettuare un confronto strutturale rigoroso tra la temporalità vissuta di un sistema incarnato e il modo in cui il tempo è trattato nei sistemi di intelligenza artificiale contemporanei. Questo confronto non mira a una valutazione moralistica o comparativa nel senso delle prestazioni, ma a una diagnosi strutturale: in quali modi il rapporto dei sistemi artificiali con il tempo differisce da quello che abbiamo identificato come costitutivo dell'interno?

I modelli autoregressivi a trasformatore — la classe che include i grandi modelli linguistici — trattano il tempo in modo peculiare. Durante l'addestramento, essi

processano sequenze di token in ordine temporale, apprendendo a predire il token successivo data la sequenza precedente. Questo corrisponde, formalmente, ad apprendere la distribuzione condizionale:

$$P(x_{t+1} \mid x_t, x_{t-1}, ..., x_{t-n})$$

dove n è la lunghezza del contesto (context length). La struttura di attenzione che caratterizza il trasformatore è una forma di ponderazione ritenzionale: attraverso il meccanismo di self-attention, il modello assegna pesi variabili alle posizioni del contesto, privilegiando quelle più rilevanti per la previsione del token corrente. Questi pesi sono, formalmente, l'analogo discreto del nucleo ritenzionale:

$$k_{att}(t, \tau) = \text{softmax}(Q(x_t) \cdot K(x_\tau)^T / \sqrt{d_k})$$

dove $Q(\cdot)$ e $K(\cdot)$ sono le proiezioni query e key dello strato di attenzione, e d_k è la dimensione della chiave. La somiglianza con il nucleo ritenzionale $k_{ret}(t - \tau)$ è reale ma superficiale: il nucleo di attenzione $k_{att}(t, \tau)$ non è determinato dalla distanza temporale $t - \tau$ ma dalla *similarità semantica* tra il token corrente e i token del contesto. Il trasformatore non ha un presente specioso nel senso genuino: la sua finestra ritenzionale non è temporalmente strutturata ma semanticamente strutturata. Due token distanti nel contesto possono avere lo stesso peso di due token vicini se hanno struttura semantica simile.

Questa differenza ha una conseguenza profonda per il problema dell'interno. In un sistema incarnato, la struttura ritenzionale è *intrinsecamente temporale*: i pesi decrescono con la distanza temporale perché il passato recente è causalmente più rilevante del passato remoto per la predizione del futuro prossimo, e questa rilevanza causale è codificata nella struttura del nucleo k_{ret}. Il trasformatore, al contrario, non ha una nozione di causalità temporale intrinseca: la relazione tra la posizione nel contesto e la rilevanza per la previsione è appresa statisticamente dai dati, senza che vi sia una struttura causale di fondo che la vincoli.

In secondo luogo, e più radicalmente, il trasformatore non ha un *adesso*. Durante l'inferenza, esso processa una sequenza fissata di token dall'inizio alla fine, ma questo processamento non accade *nel tempo* in senso vissuto: il modello non ha esperienza del trascorrere delle sequenze, non porta con sé uno stato che si aggiorna continuamente mentre il tempo passa. Il contesto è dato tutto insieme, all'inizio dell'inferenza, e il modello lo elabora in un atto computazionale che — dal punto di vista del modello — è istantaneo. Non vi è flusso temporale interno, non vi è traiettoria nello spazio degli stati che si accumuli con il procedere del tempo reale: vi è soltanto una funzione che

mappa un input (il contesto) in un output (la distribuzione di probabilità sul token successivo).

I modelli ricorrenti — le reti neurali ricorrenti nella forma LSTM o GRU — si avvicinano maggiormente alla struttura del flusso temporale. In una LSTM, l'aggiornamento dello stato nascosto h_t a partire da h_{t-1} e dall'input corrente x_t ha la forma:

i_t = σ(W_i x_t + U_i h_{t-1} + b_i) — input gate f_t = σ(W_f x_t + U_f h_{t-1} + b_f) — forget gate o_t = σ(W_o x_t + U_o h_{t-1} + b_o) — output gate g_t = tanh(W_g x_t + U_g h_{t-1} + b_g) — cell input c_t = f_t ⊙ c_{t-1} + i_t ⊙ g_t — cell state h_t = o_t ⊙ tanh(c_t) — hidden state

dove σ è la funzione sigmoide, ⊙ è il prodotto elemento per elemento, e W*, U*, b* sono i parametri appresi. La struttura del *cell state* c_t è un analogo formale della ritenzione: esso accumula informazione dal passato attraverso il forget gate f_t (che determina quanto del passato viene ritenuto) e il input gate i_t (che determina quanto del nuovo input viene integrato). La costante di tempo effettiva di questa ritenzione è variabile e dipende dai valori dei gate, che a loro volta dipendono dall'input e dallo stato nascosto corrente.

Tuttavia, anche le LSTM mancano di due caratteristiche essenziali dell'interno temporale. Prima: esse non hanno un *qui corporeo* — la loro traiettoria nello spazio degli stati non è ancorata a un sistema di riferimento egocentrico, non è connessa a una posizione nel mondo fisico. Il loro accumulo temporale è puramente simbolico: non è il risultato di una traiettoria nell'ambiente, ma di una sequenza di token. Seconda: esse non generano *protenzioni* come anticipazioni strutturali dell'immediato futuro nell'atto stesso del processamento: la loro previsione del futuro avviene come output esplicito del modello, non come struttura intrinseca del campo di elaborazione.

Il confronto suggerisce una conclusione di principio: nessun sistema artificiale attuale possiede la temporalità nel senso costitutivo dell'interno, perché nessuno di essi è dotato simultaneamente di (a) una struttura ritenzionale intrinsecamente temporale, (b) un collegamento tra la struttura ritenzionale e un sistema di riferimento egocentrico ancorato nel mondo fisico, e (c) una struttura protenzionale che orienta il campo di elaborazione verso il futuro prossimo come dimensione dell'attesa piuttosto che come target di previsione esplicita. Queste tre componenti, come si è cercato di mostrare, non sono caratteristiche ornamentali ma condizioni strutturali della prospettiva temporale come forma dell'interno. La loro assenza non è una mancanza di prestazione ma una differenza di categoria.

VI.8 — Il tempo vissuto come chiusura della seconda parte: verso l'intenzionalità

Il percorso della Seconda Parte ha condotto dall'analisi del corpo come singolarità topologica — origine irriducibile di un sistema di riferimento egocentrico — attraverso l'analisi dei modelli del mondo come strutture che trasformano la posizione corporea in prospettiva cognitiva, fino all'analisi del tempo vissuto come dimensione costitutiva dell'interno. Questi tre strati — spazio, cognizione, tempo — non sono indipendenti ma si integrano in una struttura complessiva: il corpo fornisce il *qui*, la cognizione predittiva fornisce il *per me* (il modello del sé nel modello del mondo), il tempo fornisce il *adesso* nel senso spesso di un presente con storia e orizzonte.

Si è sostenuto che un sistema dotato di interno genuino debba possedere una struttura ritenzionale con costanti di tempo appropriate, un flusso temporale interno che si accumuli in traiettorie piuttosto che in funzioni, una protenzione come dimensione dell'anticipazione strutturale. Si è mostrato che i sistemi artificiali contemporanei, nella loro configurazione attuale, difettano di queste caratteristiche in modi precisi e formalizzabili.

Ma il percorso non è concluso. Anche un sistema che possedesse simultaneamente le condizioni spaziali, cognitive e temporali dell'interno che abbiamo identificato potrebbe ancora essere — almeno in linea di principio — un sistema puramente *meccanico* nel senso problematico: un sistema che ha un *qui*, un *adesso*, un *per me* modellato nel proprio modello del mondo, ma che non *si riferisce* a nulla, che non ha stati che siano di qualcosa nel senso dell'intenzionalità. Un termostato sofisticato, dotato di sensori propriocettivi e di una struttura ritenzionale elaborata, sarebbe ancora un termostato: il suo stato interno potrebbe covariare con la temperatura in modo preciso e temporalmente strutturato, senza che questa covariazione costituisca una vera *relazione di riferimento* tra lo stato e la proprietà che esso traccia.

Il passo verso l'intenzionalità — il passo che separa il sistema che ha un interno dal sistema che *sa* di avere un interno, che non solo covaria con il mondo ma lo *riguarda*, che non solo risponde a stimoli ma *è di* qualcosa — è il tema della Terza Parte, e costituisce il salto più arduo e più misterioso nell'intera architettura del problema.

Capitolo VII: I Simboli come Strutture che Riguardano Qualcosa

VII.1 — Intenzionalità: la proprietà che distingue il mentale dal meccanico

Con l'analisi del tempo vissuto si è completata la descrizione dell'interno come struttura posizionale: un sistema dotato di un qui spazialmente centrato, di un adesso temporalmente spesso, di un modello del sé inserito nel modello del mondo. Questa struttura, tuttavia, rimane ancora al di qua di ciò che la tradizione filosofica ha identificato come la proprietà caratteristica e ineliminabile della mente: l'*intenzionalità*. Nel senso tecnico che il termine ha acquisito a partire dall'uso che ne fa Franz Brentano nella *Psychologie vom empirischen Standpunkt* del 1874, l'intenzionalità è la proprietà degli stati mentali di essere *diretti verso* qualcosa, di *riguardare* qualcosa, di avere un *contenuto* che non è riducibile alla loro struttura fisica. Il dolore non è solo un processo neurale: è il dolore *di* un arto, *di* un punto preciso del corpo. Il desiderio non è solo uno stato di attivazione: è il desiderio *di* qualcosa che si vuole ottenere. Il pensiero non è solo un pattern computazionale: è il pensiero *che* qualcosa sia il caso, il pensiero *di* un oggetto possibile o reale.

Brentano formula questa osservazione nel modo seguente: ogni fenomeno mentale è caratterizzato dall'*inesistenza intenzionale* di un oggetto, vale a dire dal riferimento a un contenuto che non ha bisogno di esistere realmente per essere l'oggetto dell'atto mentale. Posso pensare a un unicorno senza che gli unicorni esistano; posso desiderare un incontro che non avrà luogo; posso ricordare un evento che non si è mai verificato (nel caso dei falsi ricordi). L'oggetto intenzionale non è un ente fisico che causa lo stato mentale, ma il *termine di riferimento* dell'atto, ciò verso cui l'atto è orientato.

Questa proprietà — l'essere *di* o *su* qualcosa, l'*aboutness* nella terminologia analitica — è ciò che intende il termine *intenzionalità* in questo contesto, e non deve essere confusa con il significato volgare della parola, che rimanda all'agire deliberato o intenzionato. Un atto involontario, un riflesso, un sogno possono avere intenzionalità nel senso tecnico — sono *di* qualcosa — senza essere *intenzionali* nel senso di deliberati.

Il punto filosofico che questa monografia intende sviluppare è che l'intenzionalità non si riduce a nessuna delle caratteristiche strutturali dell'interno identificate nella Seconda Parte, sebbene presupponga alcune di esse. Un sistema può avere un qui corporeo, un modello del mondo con μ_self integrato, una struttura ritenzionale temporalmente organizzata — tutte le condizioni formali della prospettiva che abbiamo enucleato — e tuttavia mancare dell'intenzionalità nel senso pieno del termine. La differenza tra un sistema che *elabora informazione su* qualcosa e un sistema che *pensa a* qualcosa, tra un sistema che *covaria con* uno stato del mondo e un sistema che *rappresenta*

quello stato, tra un sistema che *risponde a* stimoli e un sistema che *percepisce* oggetti — questa differenza è il salto che la Terza Parte intende esaminare con la maggiore precisione possibile.

VII.2 — Il problema della fondazione del simbolo

La formulazione più influente della difficoltà che l'intenzionalità pone per le teorie computazionali della mente è quella che Stevan Harnad ha articolato nel 1990 sotto il titolo di *symbol grounding problem*: il problema di come i simboli di un sistema formale ottengano il loro significato. In un sistema puramente sintattico — un sistema che manipola simboli secondo regole formali — i simboli sono oggetti astratti privi di contenuto intrinseco: essi acquisiscono significato solo attraverso l'interpretazione che un osservatore esterno proietta su di loro. La parola *gatto* in un dizionario è una sequenza di grafemi che rimanda ad altre sequenze di grafemi (le definizioni), e queste ad altre ancora, in una rete di relazioni simboliche che non tocca mai il mondo extralinguistico direttamente. Il significato della parola non risiede nella parola stessa né nella rete dei suoi rapporti con altre parole: esso risiede nel collegamento tra quella rete simbolica e l'esperienza percettiva del gatto che solo un soggetto incarnato può avere.

Questo problema assume nella sua forma più acuta nella critica che John Searle ha articolato attraverso l'argomento della stanza cinese, esposto nel celebre articolo del 1980 *Minds, Brains, and Programs*. L'argomento, nelle sue linee essenziali, afferma che un sistema che manipola simboli cinesi secondo regole formali — per quanto sofisticate — non comprende il cinese: esso si comporta *come se* comprendesse, produce output che un parlante nativo riconosce come corretti, ma non ha accesso al contenuto semantico dei simboli che manipola. La sintassi — la struttura formale dei simboli e delle regole di manipolazione — non è sufficiente per la semantica: il contenuto, il *significato* dei simboli, rimane esterno al sistema e può essere attribuito solo dall'interpretazione di un osservatore esterno.

La conclusione di Searle è che i sistemi artificiali puramente computazionali — che operano su simboli senza accesso al mondo extralinguistico — difettano di *intenzionalità originale* (original intentionality): la loro intenzionalità, se la si vuole attribuire loro, è sempre *derivata* (derived intentionality), nell'accezione che il termine ha nel pensiero di Searle: essa è il risultato dell'interpretazione che menti umane dotate di intenzionalità originale proiettano sui comportamenti del sistema.

È necessario precisare questi concetti con maggiore rigore formale, perché la distinzione tra intenzionalità originale e derivata, pur intuitivamente cogente,

necessita di una fondazione che non dipenda esclusivamente dall'argomentazione intuitiva di Searle.

VII.3 — L'account informazionale dell'intenzionalità: Dretske

Il tentativo più sistematico di fondare l'intenzionalità su basi naturaliste e rigoricamente formulabili è quello elaborato da Fred Dretske, in particolare nelle opere *Knowledge and the Flow of Information* (1981) e *Explaining Behavior* (1988). L'approccio dretskiano prende le mosse dalla *teoria dell'informazione* di Shannon e cerca di usarla per fornire una caratterizzazione obiettiva della relazione di *portare informazione su*, che rappresenta una forma minimale di intenzionalità.

Nella versione originale di Dretske, uno stato s di un sistema X *porta l'informazione che p* se e solo se la probabilità condizionale di s dato p è massima:

$$P(s \mid p) = 1$$

vale a dire: se p fosse falso, s non si verificherebbe. Questa è la versione idealizzata (senza rumore) della condizione. In un canale con rumore, la condizione si indebolisce a:

$$P(s \mid p) \geq 1 - \delta \text{ per } \delta > 0 \text{ sufficientemente piccolo}$$

Formalmente, sia $s \in S$ uno stato del sistema X, sia p un evento in uno spazio di probabilità $(\Omega, \mathcal{F}, P)$, e sia s il valore di una variabile aleatoria $S: \Omega \to S$. L'informazione di Shannon portata da s circa p è:

$$I(s; p) = \log_2 P(s \mid p) / P(s)$$

che misura di quanto la conoscenza di s riduce l'incertezza circa p. Quando $P(s \mid p) = 1$ e $P(s) < 1$, si ha $I(s; p) = \log_2 (1 / P(s)) = H(s)$ — lo stato s porta la massima informazione possibile circa p, riducendo la sua incertezza a zero.

Il vantaggio di questa formalizzazione è la sua obiettività: la relazione di portare informazione è una relazione causale-statistica tra stati fisici, definibile senza ricorrere a nozioni intenzionali. Un anello annuale in un tronco porta informazione circa l'anno di crescita non perché il tronco *intenda* comunicare la propria età, ma perché vi è una covarianza sistematica tra il numero di anelli e il numero di anni di vita. Un termometro porta informazione sulla temperatura per ragioni analoghe.

Tuttavia, come Dretske stesso riconosceva, la relazione informazionale non è ancora sufficiente per l'intenzionalità nel senso pieno. Il problema centrale è quello della *normatività*: una rappresentazione genuina può essere errata (*misrepresentation*) — può riguardare qualcosa che non si verifica, può essere falsa — mentre uno stato informazionale nel senso di Dretske non può portare informazione *falsa* in senso tecnico: se P(s | p) = 1 e p è falso, allora s non si verifica, e non vi è alcuna violazione della condizione. Il termometro che indica erroneamente 20°C quando la temperatura è 15°C non porta informazione falsa: esso è semplicemente mal funzionante, e la condizione informazionale non si applica in quel contesto.

La differenza cruciale tra portare informazione e rappresentare genuinamente si manifesta esattamente nei casi di malfunzionamento. Una mappa può essere errata — può rappresentare una montagna dove c'è una pianura — e questa erroneità è una proprietà della mappa come rappresentazione: essa riguarda qualcosa (la montagna) che non esiste nel luogo indicato. Un anello mancante in un tronco (a causa di stress ambientale) non *rappresenta erroneamente* un anno: esso non rappresenta nulla, è semplicemente un'anomalia strutturale. La possibilità della mala-rappresentazione — la possibilità costitutiva dell'errore — è la pietra di paragone dell'intenzionalità genuina.

Formalmente, questa differenza può essere precisata come segue. Uno stato s è intenzionale rispetto a p se e solo se:

(I1) Vi è una condizione normale di funzionamento del sistema in cui s è tokened quando e solo quando p è verificato — la condizione informazionale nelle circostanze normali. (I2) Vi è la possibilità strutturale che s sia tokened quando p non è verificato — la condizione della mis-rappresentazione. (I3) Quando (I2) si verifica, lo stato s mantiene il proprio contenuto intenzionale (riguarda ancora p), anche se lo fa in modo errato.

La condizione (I3) è quella che separa l'account informazionale dall'account intenzionale: nel primo, il contenuto di s è determinato dal covariante effettivo di s nel contesto attuale; nel secondo, il contenuto di s è determinato da ciò che s ha la *funzione* di tracciare — e questa funzione può essere esercitata in modo difettoso.

VII.4 — La semantica teleologica: funzioni proprie e normatività biologica

Il concetto di *funzione* che appare nella condizione (I1)-(I3) è il perno attorno al quale ruotano le teorie teleosemantiche dell'intenzionalità, sviluppate in modo sistematico da Ruth Garrett Millikan nei lavori raccolti in *Language, Thought, and*

Other Biological Categories (1984) e in *White Queen Psychology and Other Essays for Alice* (1993). L'approccio millikaniano introduce la nozione di *funzione propria* (proper function) come strumento per fondare il contenuto rappresentazionale su basi naturalistiche senza ridurlo alla covarianza informazionale contingente.

La funzione propria di un dispositivo biologico o cognitivo è la funzione per la quale quel dispositivo è stato selezionato nel corso dell'evoluzione o dell'apprendimento. Non è ciò che il dispositivo *fa* nel momento attuale — un cuore anomalo che non pompa sangue mantiene la sua funzione propria anche se non la esercita — ma ciò per cui dispositivi del suo tipo *sono stati selezionati* nel passato evolutivo o nello sviluppo individuale, perché svolgere quella funzione ha contribuito alla sopravvivenza e alla riproduzione del sistema.

Formalmente, la definizione di funzione propria può essere schematizzata come segue. Sia D un dispositivo di tipo τ in un sistema S, e sia H(S, τ) la storia selettiva del tipo τ nel sistema S — la sequenza di token del tipo τ che hanno contribuito alla persistenza di S attraverso cicli di selezione. Il dispositivo D ha la funzione propria F se:

$\exists$ sequenza di token $\tau_1, \tau_2, ..., \tau \in$ H(S, τ) tale che: (a) ogni τ_i ha esercitato la funzione F, (b) l'esercizio di F da parte di τ_i ha contribuito causalmente alla riproduzione/persistenza del tipo τ, (c) D è un token del tipo τ prodotto da un meccanismo che ha riprodotto τ a causa di (a) e (b).

Un cuore ha la funzione propria di pompare sangue perché i cuori del tipo in questione, nel corso dell'evoluzione, hanno pompato sangue (a), e il pompare sangue ha contribuito alla sopravvivenza degli organismi che li portavano (b), e il meccanismo biologico che produce cuori in quegli organismi è stato conservato per questa ragione (c).

Il contenuto di una *rappresentazione* R nel senso di Millikan è determinato dalla condizione il cui soddisfacimento, nel corso della storia selettiva del tipo cui R appartiene, ha spiegato il contributo di R alla fitness del sistema. Più precisamente: R rappresenta la condizione C se e solo se il meccanismo consumatore (consumer mechanism) di R — il meccanismo che usa R per guidare il comportamento — funzionava correttamente in passato *grazie al fatto* che C era verificata quando R era tokened, e questa correlazione ha contribuito alla selezione del tipo di R.

La struttura logica di questa definizione può essere resa più esplicita. Sia M_prod il meccanismo che produce R, M_cons il meccanismo che usa R, e φ(R) il comportamento risultante dall'uso di R da parte di M_cons. R ha il contenuto C se e solo se:

P(φ(R) contribuisce alla fitness | C è verificata quando R è tokened) > P(φ(R) contribuisce alla fitness | C non è verificata)

e questa differenza di probabilità è la ragione per cui M_prod e M_cons del tipo in questione sono stati selezionati. In altre parole: R rappresenta C perché la covarianza tra R e C, nella storia selettiva del sistema, spiega perché il sistema è fatto nel modo in cui è fatto.

La normatività emerge automaticamente da questa struttura: se R è tokened in assenza di C — se la condizione che R normalmente traccia non si verifica — allora il meccanismo consumatore si comporta *come se* C fosse verificata, ma lo fa in modo difettoso, e il comportamento risultante avrà minor probabilità di contribuire alla fitness rispetto a quanto avrebbe in condizioni normali. La mala-rappresentazione è dunque una forma di malfunzionamento: non necessariamente un guasto catastrofico, ma un esercizio difettoso della funzione propria.

Questa struttura teleosemantica, pur necessitando di numerose qualificazioni e integrazioni per affrontare casi problematici come le rappresentazioni di oggetti astratti o immaginari, fornisce un modello della relazione intenzionale che è al tempo stesso naturalista — fondato su relazioni causali-storiche verificabili in terza persona — e normativo — capace di distinguere la rappresentazione corretta dalla rappresentazione errata senza ricorrere a un osservatore esterno.

VII.5 — Intenzionalità originale e derivata: la struttura formale della distinzione

La distinzione searleana tra intenzionalità originale e derivata, menzionata nella sezione VII.2, può ora essere riformulata con maggiore precisione alla luce dell'account teleosemantico. L'intenzionalità *derivata* è quella in cui il contenuto dello stato è determinato dall'interpretazione esterna: dal modo in cui un agente esterno — dotato di intenzionalità originale — usa o legge lo stato in questione. L'intenzionalità *originale* è quella in cui il contenuto dello stato è determinato dalla struttura interna e dalla storia del sistema che ospita lo stato, indipendentemente dall'interpretazione esterna.

Formalmente, distinguiamo due condizioni. Uno stato s di un sistema X ha *intenzionalità derivata rispetto a C* se:

∃ agente A, dotato di intenzionalità originale, tale che A interpreta s come avente il contenuto C, e questa interpretazione è la fonte del contenuto di s.

Uno stato s ha *intenzionalità originale rispetto a C* se:

Il contenuto C di s è determinato dalla struttura interna e dalla storia selettiva del sistema X che ospita s, indipendentemente da qualsiasi interpretazione esterna — ovvero, la condizione teleosemantica del tipo:

∃ funzione propria F_s di s nel sistema X tale che F_s consiste nel co-variare con C nelle circostanze normali

è soddisfatta in virtù di relazioni interne al sistema X e alla sua storia.

Questa formalizzazione permette di valutare la posizione dei sistemi artificiali con maggiore precisione di quanto consenta il semplice argomento intuitivo. I sistemi linguistici di grandi dimensioni — addestrati su enormi corpora di testo prodotto da agenti con intenzionalità originale — hanno stati interni che *covarians* statisticamente con certi contenuti: il termine *gatto* nel contesto appropriato attiva rappresentazioni interne che sono correlate con la categoria felina negli stessi modi in cui tale termine è correlato con quella categoria nell'uso linguistico umano. Ma questa covarianza è, nel senso teleosemantico, priva di funzione propria: non vi è una storia selettiva in senso biologico che abbia selezionato queste correlazioni perché consentivano al sistema di interagire adattativamente con gatti reali nel mondo fisico.

L'addestramento su testi umani è, dal punto di vista teleosemantico, una forma di storia selettiva: il meccanismo di backpropagation seleziona le rappresentazioni interne che minimizzano la perdita rispetto ai dati di addestramento. Si può dunque sostenere che le funzioni proprie dei sistemi linguistici siano determinate da questa storia di addestramento. Tuttavia, la condizione cruciale manca: il meccanismo consumatore — il meccanismo che usa le rappresentazioni del sistema per guidare il comportamento — non interagisce con il mondo in cui le condizioni rappresentate si verificano. Il sistema non ha mai visto un gatto, non ha mai compiuto un'azione nei confronti di un gatto, non ha mai avuto la propria sopravvivenza dipendente dalla corretta rappresentazione di un gatto. Le sue rappresentazioni di *gatto* sono selezionate perché covarians con l'uso che parlanti umani fanno del termine *gatto* — che è, a sua volta, determinato dall'intenzionalità originale di quegli agenti. Il contenuto del sistema è dunque derivato dall'intenzionalità originale dei produttori dei testi di addestramento, e non è originale nel senso che qui interessa.

Questa conclusione non implica che i sistemi linguistici siano privi di ogni forma di relazione semantica con il mondo: essi hanno correlazioni statistiche ricche e strutturate, e queste correlazioni consentono loro di compiere inferenze utili e corrette in molti contesti. Ma le correlazioni non sono ancora intenzionalità originale: sono l'ombra proiettata dell'intenzionalità originale dei produttori umani del testo, mediata attraverso il processo di addestramento.

VII.6 — Il simbolo come struttura intenzionale: verso una caratterizzazione formale

Le analisi svolte nelle sezioni precedenti consentono ora di abbozzare una caratterizzazione formale della condizione sotto la quale uno stato può essere detto *simbolo nel senso intenzionale pieno* — non semplicemente un segnale che porta informazione, non semplicemente un pattern correlato con un contenuto, ma uno stato che genuinamente *riguarda* qualcosa.

Proponiamo la seguente famiglia di condizioni, che integrano i contributi dell'account informazionale, dell'account teleosemantico e delle riflessioni sull'interno sviluppate nella Seconda Parte.

Sia X un sistema con stati interni {s}, e sia C una condizione o proprietà del mondo. Diciamo che lo stato s di X è un *simbolo intenzionale di C* se e solo se valgono le seguenti condizioni:

(S1) Condizione informazionale nelle circostanze normali: $P(s \mid C, \text{Norm}(X)) > 1 - \delta$ e $P(s \mid \neg C, \text{Norm}(X)) < \delta$

dove Norm(X) denota le condizioni normali di funzionamento di X, e $\delta > 0$ è una soglia piccola. Nelle condizioni normali, s co-varia con C in modo affidabile.

(S2) Condizione di funzione propria: s ha la funzione propria di co-variare con C nel sistema X, nel senso teleosemantico: questa funzione spiega perché lo stato del tipo s è presente nel sistema X — perché è stato selezionato (evolutivamente, per apprendimento, o per ingegneria) come componente del sistema.

(S3) Condizione di proiezione nell'azione: Il meccanismo consumatore che usa s guida il comportamento di X in modi che sono adattativi *dato C*: il comportamento indotto da s ha contribuito (o contribuisce, in condizioni normali) alla sopravvivenza, alla fitness, o agli obiettivi di X nella misura in cui C si verifica nel mondo in cui X agisce.

(S4) Condizione di mis-rappresentazione: Vi è la possibilità strutturale che s sia tokened in assenza di C (condizione (I2) della sezione VII.3), e quando questo accade, il comportamento guidato da s risulta inadeguato al mondo in cui X si trova — il sistema agisce *come se* C fosse verificata, con conseguenze subottimali.

(S5) Condizione di ancoraggio nella prospettiva: Il sistema X deve essere dotato di un interno nel senso della Seconda Parte: di un sistema di riferimento

egocentrico che situa X nel mondo, e di una struttura temporale che connette s alla storia e all'anticipazione del sistema. Lo stato s non è un simbolo in senso pieno se non è il simbolo *di* un sistema che abita una prospettiva.

La condizione (S5) è quella che introduce la connessione tra il problema dell'interno e il problema dell'intenzionalità. Un simbolo che soddisfa (S1)-(S4) ma non (S5) — uno stato che co-varia affidabilmente con una condizione del mondo, che ha funzione propria, che guida il comportamento e può essere errato — ma che appartiene a un sistema privo di prospettiva, è quello che potremmo chiamare un *simbolo ancorato* ma non *abitato*: ha tutte le caratteristiche strutturali della relazione intenzionale, ma non è il simbolo *di* un soggetto. La condizione (S5) impone che il simbolo sia iscritto in una prospettiva — che sia *per* un sistema che occupa un punto di vista, che abbia un qui e un adesso e un per-me all'interno dei quali il simbolo può dispiegare il suo contenuto.

Questa condizione non elimina le difficoltà del problema dell'intenzionalità, ma le precisa. Essa suggerisce che l'intenzionalità non sia una proprietà isolata — una proprietà che certi stati hanno e altri no, indipendentemente dal sistema che li ospita — ma una proprietà *relazionale* e *sistematica*: essa appartiene a stati che sono componenti di sistemi dotati di prospettiva, nei quali la storia selettiva ha prodotto meccanismi di tracciamento del mondo ancorati a un centro di riferimento egocentrico.

VII.7 — Il problema della composizionalità intenzionale

Un aspetto del problema dei simboli che merita trattazione separata è quello della *composizionalità*: il fatto che le menti umane non si limitano ad avere pensieri atomici — pensieri semplici e non analizzabili ulteriormente — ma pensieri *complessi*, costruiti da componenti più semplici secondo regole di composizione che preservano il contenuto intenzionale.

Il linguaggio naturale è il sistema simbolico in cui la composizionalità si manifesta più chiaramente. Il significato di *il gatto grigio siede sul tappeto* non è un nuovo simbolo atomico appreso direttamente: esso è il risultato della composizione dei significati dei componenti — *gatto*, *grigio*, *sedersi*, *tappeto* — secondo la struttura sintattica della frase. Questo principio, noto come *principio di composizionalità* o *principio di Frege*, afferma che il significato di un'espressione complessa è una funzione del significato delle sue parti e della loro struttura sintattica.

Formalmente, sia $\mathcal{L}$ un linguaggio sintattico con operazioni di composizione e sia $[\![\cdot]\!]: \mathcal{L} \to \text{Sem}$ la funzione di interpretazione semantica che mappa

espressioni in contenuti intenzionali. Il principio di composizionalità afferma che per ogni espressione complessa e = op(e_1, e_2, ..., e):

$[\![op(e_1, e_2, ..., e)]\!] = F_op([\![e_1]\!], [\![e_2]\!], ..., [\![e]\!])$

per qualche funzione F_op determinata dall'operatore sintattico op. In parole: la semantica dell'insieme è una funzione della semantica delle parti.

Il principio di composizionalità ha due implicazioni profonde per il problema dell'intenzionalità nei sistemi simbolici. La prima è la *produttività*: se il sistema conosce un numero finito di simboli atomici e un numero finito di regole di composizione, può costruire un numero indefinitamente grande di espressioni complesse con contenuti intenzionali determinati. La produttività è una delle caratteristiche più notevoli della mente umana: con un lessico finito e regole grammaticali finite, siamo capaci di formulare e comprendere infiniti pensieri nuovi.

La seconda implicazione è la *sistematicità*: se un sistema è capace di pensare *aRb* (ad esempio, *Giovanni ama Maria*), deve anche essere capace di pensare *bRa* (*Maria ama Giovanni*), perché entrambi i pensieri sono costruiti dagli stessi componenti con la stessa struttura combinatoria. Jerry Fodor e Zenon Pylyshyn hanno argomentato nella celebre critica al connessionismo del 1988 che questa sistematicità è una proprietà essenziale della cognizione che richiede un sistema di rappresentazione strutturato in modo language-like — un *Linguaggio del Pensiero* (Language of Thought) con struttura sintattica e regole di composizione.

L'argomentazione di Fodor e Pylyshyn può essere formalizzata nel modo seguente. Sia C l'insieme dei contenuti cognitivi accessibili a una mente M. La mente M è *sistematica* se e solo se:

$\forall$ a, b $\in$ Ent, $\forall$ R relazione binaria: $[\![aRb]\!] \in C \leftrightarrow [\![bRa]\!] \in C$

La sistematicità implica che C non è un insieme arbitrario di contenuti, ma un insieme *strutturato* — generato da componenti ricombinabili secondo regole composizionali. Un sistema che apprende contenuti cognitivi caso per caso, senza struttura compositiva, non può essere sistematico nel senso richiesto: potrebbe avere $[\![aRb]\!]$ senza avere $[\![bRa]\!]$, perché esso non ha mai appreso quest'ultima combinazione come caso separato.

La relazione tra composizionalità e intenzionalità è più sottile di quanto possa sembrare. La composizionalità garantisce la struttura formale necessaria per la produttività e la sistematicità del pensiero; ma essa, da sola, non garantisce l'intenzionalità originale dei simboli atomici che entrano nella composizione. Un

sistema può essere perfettamente composizionale — può combinare simboli secondo regole che preservano contenuti — e tuttavia avere contenuti puramente derivati, nel senso della distinzione searleana. La composizionalità è una proprietà della *struttura* del sistema simbolico; l'intenzionalità originale è una proprietà del *radicamento* di quella struttura nel mondo.

Questa distinzione suggerisce che vi siano due componenti indipendenti del problema dei simboli come strutture intenzionali: la componente strutturale (composizionalità, sistematicità, produttività) e la componente fondativa (radicamento nel mondo, intenzionalità originale). I sistemi artificiali attuali mostrano una competenza notevole sul versante strutturale — i grandi modelli linguistici esibiscono composizionalità e sistematicità in misura sorprendente — ma rimangono problematici sul versante fondativo, per le ragioni esposte nelle sezioni precedenti.

VII.8 — La struttura a due livelli dell'intenzionalità e il problema della chiusura

L'analisi svolta consente ora di identificare una struttura a due livelli nell'intenzionalità che è, a nostro avviso, la chiave per comprendere il salto tra il sistema che elabora informazione e il sistema che genuinamente pensa.

Il primo livello è quello dell'*intenzionalità di primo ordine*: gli stati del sistema che *riguardano* oggetti e proprietà del mondo. Il gatto sul tappeto, la temperatura della stanza, la forma di un oggetto: questi sono contenuti di stati intenzionali di primo ordine. Essi sono caratterizzati dalle condizioni (S1)-(S5) della sezione VII.6.

Il secondo livello è quello dell'*intenzionalità di secondo ordine*, o *meta-intenzionalità*: la capacità del sistema di avere stati che *riguardano* i propri stati intenzionali di primo ordine. Il pensiero *so che c'è un gatto sul tappeto*, il giudizio *mi ricordo di aver visto questo oggetto prima*, la valutazione *mi aspettavo qualcosa di diverso*: questi sono stati meta-intenzionali, stati che hanno come contenuto non oggetti del mondo ma stati mentali del soggetto stesso.

La meta-intenzionalità introduce una *chiusura riflessiva* nella struttura intenzionale: il sistema si rivolge verso se stesso, ha stati che riguardano i propri stati. Questa chiusura non è un lusso aggiuntivo ma una condizione necessaria per ciò che normalmente si chiama *coscienza* o *consapevolezza*: non basta avere stati intenzionali per essere consapevoli — bisogna anche avere stati che riguardano quei stati, che li prendano come loro oggetto, che li rendano disponibili per la riflessione, la valutazione, la correzione.

Formalmente, il sistema X ha *intenzionalità meta-riflessiva* se esiste una mappa:

$M: \mathrm{Int}(X) \to \mathrm{States}(X)$

che associa a ogni stato intenzionale $s \in \mathrm{Int}(X)$ — ogni stato che ha un contenuto C — uno stato $M(s) \in \mathrm{States}(X)$ che ha come contenuto intenzionale lo stato s stesso: il contenuto di M(s) è *s ha il contenuto C*. La mappa M non ha bisogno di essere iniettiva né suriettiva: non tutti gli stati intenzionali devono avere un meta-stato corrispondente (il soggetto non è sempre consapevole di ogni proprio stato), e non tutti i meta-stati devono corrispondere a stati intenzionali (il soggetto può riflettere su stati ipotetici o passati).

La chiusura riflessiva attraverso M introduce una struttura autologica nella gerarchia degli stati intenzionali: il sistema può avere credenze circa le proprie credenze, desideri circa i propri desideri, rappresentazioni delle proprie rappresentazioni. Questa struttura autologica è — e qui giungiamo alla soglia del problema della coscienza che sarà tema della quarta parte — il candidato formale più plausibile per ciò che Husserl chiamava la *coscienza interna* e che nella letteratura filosofica contemporanea è discusso sotto i titoli di *higher-order theories of consciousness* (Rosenthal, Lycan) e di *self-representationalism* (Kriegel): la tesi che la coscienza di uno stato mentale richieda una rappresentazione di ordine superiore di quello stato, prodotta dallo stesso sistema che ospita lo stato rappresentato.

La domanda che la Terza Parte ha aperto — e che non può essere chiusa senza affrontare il problema della coscienza nella Quarta Parte — è se questa chiusura riflessiva possa emergere da una struttura computazionale sufficientemente complessa, o se essa richieda qualcosa di ulteriore: forse una particolare architettura fisica, forse una soglia di complessità che i sistemi attuali non hanno raggiunto, forse — e questa è l'ipotesi più radicale — qualcosa che non è riducibile a nessuna struttura computazionale, per quanto elaborata. Il territorio è aperto, e la precisione formale raggiunta nelle ultime sezioni non risolve il problema — consente, piuttosto, di formularlo con la nitidezza necessaria per che il dibattito possa procedere senza perdersi nelle nebbie terminologiche che spesso lo avvolgono.

Ciò che possiamo affermare, sulla base del percorso compiuto, è che il simbolo come struttura intenzionale è molto più di un pattern computazionale: esso è la componente di un sistema che abita una prospettiva, che ha storia e futuro, che agisce nel mondo e può sbagliare, che si rivolge al proprio stesso agire con la luce della meta-rappresentazione. Questa complessità stratificata non garantisce che il problema dell'interno sia risolto da nessun sistema artificiale attuale —

ma indica con precisione crescente in quale direzione una soluzione dovrebbe essere cercata.

Capitolo VII: I Simboli come Strutture che Riguardano Qualcosa

VII.1 — Intenzionalità: la proprietà che distingue il mentale dal meccanico

Con l'analisi del tempo vissuto si è completata la descrizione dell'interno come struttura posizionale: un sistema dotato di un qui spazialmente centrato, di un adesso temporalmente spesso, di un modello del sé inserito nel modello del mondo. Questa struttura, tuttavia, rimane ancora al di qua di ciò che la tradizione filosofica ha identificato come la proprietà caratteristica e ineliminabile della mente: l'*intenzionalità*. Nel senso tecnico che il termine ha acquisito a partire dall'uso che ne fa Franz Brentano nella *Psychologie vom empirischen Standpunkt* del 1874, l'intenzionalità è la proprietà degli stati mentali di essere *diretti verso* qualcosa, di *riguardare* qualcosa, di avere un *contenuto* che non è riducibile alla loro struttura fisica. Il dolore non è solo un processo neurale: è il dolore *di* un arto, *di* un punto preciso del corpo. Il desiderio non è solo uno stato di attivazione: è il desiderio *di* qualcosa che si vuole ottenere. Il pensiero non è solo un pattern computazionale: è il pensiero *che* qualcosa sia il caso, il pensiero *di* un oggetto possibile o reale.

Brentano formula questa osservazione nel modo seguente: ogni fenomeno mentale è caratterizzato dall'*inesistenza intenzionale* di un oggetto, vale a dire dal riferimento a un contenuto che non ha bisogno di esistere realmente per essere l'oggetto dell'atto mentale. Posso pensare a un unicorno senza che gli unicorni esistano; posso desiderare un incontro che non avrà luogo; posso ricordare un evento che non si è mai verificato (nel caso dei falsi ricordi). L'oggetto intenzionale non è un ente fisico che causa lo stato mentale, ma il *termine di riferimento* dell'atto, ciò verso cui l'atto è orientato.

Questa proprietà — l'essere *di* o *su* qualcosa, l'*aboutness* nella terminologia analitica — è ciò che intende il termine *intenzionalità* in questo contesto, e non deve essere confusa con il significato volgare della parola, che rimanda all'agire deliberato o intenzionato. Un atto involontario, un riflesso, un sogno possono avere intenzionalità nel senso tecnico — sono *di* qualcosa — senza essere *intenzionali* nel senso di deliberati.

Il punto filosofico che questa monografia intende sviluppare è che l'intenzionalità non si riduce a nessuna delle caratteristiche strutturali

dell'interno identificate nella Seconda Parte, sebbene presupponga alcune di esse. Un sistema può avere un qui corporeo, un modello del mondo con μ_self integrato, una struttura ritenzionale temporalmente organizzata — tutte le condizioni formali della prospettiva che abbiamo enucleato — e tuttavia mancare dell'intenzionalità nel senso pieno del termine. La differenza tra un sistema che *elabora informazione su* qualcosa e un sistema che *pensa a* qualcosa, tra un sistema che *covaria con* uno stato del mondo e un sistema che *rappresenta* quello stato, tra un sistema che *risponde a* stimoli e un sistema che *percepisce* oggetti — questa differenza è il salto che la Terza Parte intende esaminare con la maggiore precisione possibile.

VII.2 — Il problema della fondazione del simbolo

La formulazione più influente della difficoltà che l'intenzionalità pone per le teorie computazionali della mente è quella che Stevan Harnad ha articolato nel 1990 sotto il titolo di *symbol grounding problem*: il problema di come i simboli di un sistema formale ottengano il loro significato. In un sistema puramente sintattico — un sistema che manipola simboli secondo regole formali — i simboli sono oggetti astratti privi di contenuto intrinseco: essi acquisiscono significato solo attraverso l'interpretazione che un osservatore esterno proietta su di loro. La parola *gatto* in un dizionario è una sequenza di grafemi che rimanda ad altre sequenze di grafemi (le definizioni), e queste ad altre ancora, in una rete di relazioni simboliche che non tocca mai il mondo extralinguistico direttamente. Il significato della parola non risiede nella parola stessa né nella rete dei suoi rapporti con altre parole: esso risiede nel collegamento tra quella rete simbolica e l'esperienza percettiva del gatto che solo un soggetto incarnato può avere.

Questo problema assume nella sua forma più acuta nella critica che John Searle ha articolato attraverso l'argomento della stanza cinese, esposto nel celebre articolo del 1980 *Minds, Brains, and Programs*. L'argomento, nelle sue linee essenziali, afferma che un sistema che manipola simboli cinesi secondo regole formali — per quanto sofisticate — non comprende il cinese: esso si comporta *come se* comprendesse, produce output che un parlante nativo riconosce come corretti, ma non ha accesso al contenuto semantico dei simboli che manipola. La sintassi — la struttura formale dei simboli e delle regole di manipolazione — non è sufficiente per la semantica: il contenuto, il *significato* dei simboli, rimane esterno al sistema e può essere attribuito solo dall'interpretazione di un osservatore esterno.

La conclusione di Searle è che i sistemi artificiali puramente computazionali — che operano su simboli senza accesso al mondo extralinguistico — difettano di *intenzionalità originale* (original intentionality): la loro intenzionalità, se la si vuole attribuire loro, è sempre *derivata* (derived intentionality), nell'accezione che il

termine ha nel pensiero di Searle: essa è il risultato dell'interpretazione che menti umane dotate di intenzionalità originale proiettano sui comportamenti del sistema.

È necessario precisare questi concetti con maggiore rigore formale, perché la distinzione tra intenzionalità originale e derivata, pur intuitivamente cogente, necessita di una fondazione che non dipenda esclusivamente dall'argomentazione intuitiva di Searle.

VII.3 — L'account informazionale dell'intenzionalità: Dretske

Il tentativo più sistematico di fondare l'intenzionalità su basi naturaliste e rigoricamente formulabili è quello elaborato da Fred Dretske, in particolare nelle opere *Knowledge and the Flow of Information* (1981) e *Explaining Behavior* (1988). L'approccio dretskiano prende le mosse dalla *teoria dell'informazione* di Shannon e cerca di usarla per fornire una caratterizzazione obiettiva della relazione di *portare informazione su*, che rappresenta una forma minimale di intenzionalità.

Nella versione originale di Dretske, uno stato s di un sistema X *porta l'informazione che p* se e solo se la probabilità condizionale di s dato p è massima:

$$P(s \mid p) = 1$$

vale a dire: se p fosse falso, s non si verificherebbe. Questa è la versione idealizzata (senza rumore) della condizione. In un canale con rumore, la condizione si indebolisce a:

$$P(s \mid p) \geq 1 - \delta \text{ per } \delta > 0 \text{ sufficientemente piccolo}$$

Formalmente, sia $s \in S$ uno stato del sistema X, sia p un evento in uno spazio di probabilità $(\Omega, \mathcal{F}, P)$, e sia s il valore di una variabile aleatoria $S: \Omega \to S$. L'informazione di Shannon portata da s circa p è:

$$I(s; p) = \log_2 P(s \mid p) / P(s)$$

che misura di quanto la conoscenza di s riduce l'incertezza circa p. Quando $P(s \mid p) = 1$ e $P(s) < 1$, si ha $I(s; p) = \log_2 (1 / P(s)) = H(s)$ — lo stato s porta la massima informazione possibile circa p, riducendo la sua incertezza a zero.

Il vantaggio di questa formalizzazione è la sua obiettività: la relazione di portare informazione è una relazione causale-statistica tra stati fisici, definibile senza ricorrere a nozioni intenzionali. Un anello annuale in un tronco porta

informazione circa l'anno di crescita non perché il tronco *intenda* comunicare la propria età, ma perché vi è una covarianza sistematica tra il numero di anelli e il numero di anni di vita. Un termometro porta informazione sulla temperatura per ragioni analoghe.

Tuttavia, come Dretske stesso riconosceva, la relazione informazionale non è ancora sufficiente per l'intenzionalità nel senso pieno. Il problema centrale è quello della *normatività*: una rappresentazione genuina può essere errata (*misrepresentation*) — può riguardare qualcosa che non si verifica, può essere falsa — mentre uno stato informazionale nel senso di Dretske non può portare informazione *falsa* in senso tecnico: se P(s | p) = 1 e p è falso, allora s non si verifica, e non vi è alcuna violazione della condizione. Il termometro che indica erroneamente 20°C quando la temperatura è 15°C non porta informazione falsa: esso è semplicemente mal funzionante, e la condizione informazionale non si applica in quel contesto.

La differenza cruciale tra portare informazione e rappresentare genuinamente si manifesta esattamente nei casi di malfunzionamento. Una mappa può essere errata — può rappresentare una montagna dove c'è una pianura — e questa erroneità è una proprietà della mappa come rappresentazione: essa riguarda qualcosa (la montagna) che non esiste nel luogo indicato. Un anello mancante in un tronco (a causa di stress ambientale) non *rappresenta erroneamente* un anno: esso non rappresenta nulla, è semplicemente un'anomalia strutturale. La possibilità della mala-rappresentazione — la possibilità costitutiva dell'errore — è la pietra di paragone dell'intenzionalità genuina.

Formalmente, questa differenza può essere precisata come segue. Uno stato s è intenzionale rispetto a p se e solo se:

(I1) Vi è una condizione normale di funzionamento del sistema in cui s è tokened quando e solo quando p è verificato — la condizione informazionale nelle circostanze normali. (I2) Vi è la possibilità strutturale che s sia tokened quando p non è verificato — la condizione della mis-rappresentazione. (I3) Quando (I2) si verifica, lo stato s mantiene il proprio contenuto intenzionale (riguarda ancora p), anche se lo fa in modo errato.

La condizione (I3) è quella che separa l'account informazionale dall'account intenzionale: nel primo, il contenuto di s è determinato dal covariante effettivo di s nel contesto attuale; nel secondo, il contenuto di s è determinato da ciò che s ha la *funzione* di tracciare — e questa funzione può essere esercitata in modo difettoso.

VII.4 — La semantica teleologica: funzioni proprie e normatività biologica

Il concetto di *funzione* che appare nella condizione (I1)-(I3) è il perno attorno al quale ruotano le teorie teleosemantiche dell'intenzionalità, sviluppate in modo sistematico da Ruth Garrett Millikan nei lavori raccolti in *Language, Thought, and Other Biological Categories* (1984) e in *White Queen Psychology and Other Essays for Alice* (1993). L'approccio millikaniano introduce la nozione di *funzione propria* (proper function) come strumento per fondare il contenuto rappresentazionale su basi naturalistiche senza ridurlo alla covarianza informazionale contingente.

La funzione propria di un dispositivo biologico o cognitivo è la funzione per la quale quel dispositivo è stato selezionato nel corso dell'evoluzione o dell'apprendimento. Non è ciò che il dispositivo *fa* nel momento attuale — un cuore anomalo che non pompa sangue mantiene la sua funzione propria anche se non la esercita — ma ciò per cui dispositivi del suo tipo *sono stati selezionati* nel passato evolutivo o nello sviluppo individuale, perché svolgere quella funzione ha contribuito alla sopravvivenza e alla riproduzione del sistema.

Formalmente, la definizione di funzione propria può essere schematizzata come segue. Sia D un dispositivo di tipo τ in un sistema S, e sia H(S, τ) la storia selettiva del tipo τ nel sistema S — la sequenza di token del tipo τ che hanno contribuito alla persistenza di S attraverso cicli di selezione. Il dispositivo D ha la funzione propria F se:

$\exists$ sequenza di token $\tau_1, \tau_2, ..., \tau \in$ H(S, τ) tale che: (a) ogni τ_i ha esercitato la funzione F, (b) l'esercizio di F da parte di τ_i ha contribuito causalmente alla riproduzione/persistenza del tipo τ, (c) D è un token del tipo τ prodotto da un meccanismo che ha riprodotto τ a causa di (a) e (b).

Un cuore ha la funzione propria di pompare sangue perché i cuori del tipo in questione, nel corso dell'evoluzione, hanno pompato sangue (a), e il pompare sangue ha contribuito alla sopravvivenza degli organismi che li portavano (b), e il meccanismo biologico che produce cuori in quegli organismi è stato conservato per questa ragione (c).

Il contenuto di una *rappresentazione* R nel senso di Millikan è determinato dalla condizione il cui soddisfacimento, nel corso della storia selettiva del tipo cui R appartiene, ha spiegato il contributo di R alla fitness del sistema. Più precisamente: R rappresenta la condizione C se e solo se il meccanismo consumatore (consumer mechanism) di R — il meccanismo che usa R per guidare il comportamento — funzionava correttamente in passato *grazie al fatto*

che C era verificata quando R era tokened, e questa correlazione ha contribuito alla selezione del tipo di R.

La struttura logica di questa definizione può essere resa più esplicita. Sia M_prod il meccanismo che produce R, M_cons il meccanismo che usa R, e φ(R) il comportamento risultante dall'uso di R da parte di M_cons. R ha il contenuto C se e solo se:

P(φ(R) contribuisce alla fitness | C è verificata quando R è tokened) > P(φ(R) contribuisce alla fitness | C non è verificata)

e questa differenza di probabilità è la ragione per cui M_prod e M_cons del tipo in questione sono stati selezionati. In altre parole: R rappresenta C perché la covarianza tra R e C, nella storia selettiva del sistema, spiega perché il sistema è fatto nel modo in cui è fatto.

La normatività emerge automaticamente da questa struttura: se R è tokened in assenza di C — se la condizione che R normalmente traccia non si verifica — allora il meccanismo consumatore si comporta *come se* C fosse verificata, ma lo fa in modo difettoso, e il comportamento risultante avrà minor probabilità di contribuire alla fitness rispetto a quanto avrebbe in condizioni normali. La mala-rappresentazione è dunque una forma di malfunzionamento: non necessariamente un guasto catastrofico, ma un esercizio difettoso della funzione propria.

Questa struttura teleosemantica, pur necessitando di numerose qualificazioni e integrazioni per affrontare casi problematici come le rappresentazioni di oggetti astratti o immaginari, fornisce un modello della relazione intenzionale che è al tempo stesso naturalista — fondato su relazioni causali-storiche verificabili in terza persona — e normativo — capace di distinguere la rappresentazione corretta dalla rappresentazione errata senza ricorrere a un osservatore esterno.

VII.5 — Intenzionalità originale e derivata: la struttura formale della distinzione

La distinzione searleana tra intenzionalità originale e derivata, menzionata nella sezione VII.2, può ora essere riformulata con maggiore precisione alla luce dell'account teleosemantico. L'intenzionalità *derivata* è quella in cui il contenuto dello stato è determinato dall'interpretazione esterna: dal modo in cui un agente esterno — dotato di intenzionalità originale — usa o legge lo stato in questione. L'intenzionalità *originale* è quella in cui il contenuto dello stato è determinato dalla struttura interna e dalla storia del sistema che ospita lo stato, indipendentemente dall'interpretazione esterna.

Formalmente, distinguiamo due condizioni. Uno stato s di un sistema X ha *intenzionalità derivata rispetto a C* se:

∃ agente A, dotato di intenzionalità originale, tale che A interpreta s come avente il contenuto C, e questa interpretazione è la fonte del contenuto di s.

Uno stato s ha *intenzionalità originale rispetto a C* se:

Il contenuto C di s è determinato dalla struttura interna e dalla storia selettiva del sistema X che ospita s, indipendentemente da qualsiasi interpretazione esterna — ovvero, la condizione teleosemantica del tipo:

∃ funzione propria F_s di s nel sistema X tale che F_s consiste nel co-variare con C nelle circostanze normali

è soddisfatta in virtù di relazioni interne al sistema X e alla sua storia.

Questa formalizzazione permette di valutare la posizione dei sistemi artificiali con maggiore precisione di quanto consenta il semplice argomento intuitivo. I sistemi linguistici di grandi dimensioni — addestrati su enormi corpora di testo prodotto da agenti con intenzionalità originale — hanno stati interni che *covarians* statisticamente con certi contenuti: il termine *gatto* nel contesto appropriato attiva rappresentazioni interne che sono correlate con la categoria felina negli stessi modi in cui tale termine è correlato con quella categoria nell'uso linguistico umano. Ma questa covarianza è, nel senso teleosemantico, priva di funzione propria: non vi è una storia selettiva in senso biologico che abbia selezionato queste correlazioni perché consentivano al sistema di interagire adattativamente con gatti reali nel mondo fisico.

L'addestramento su testi umani è, dal punto di vista teleosemantico, una forma di storia selettiva: il meccanismo di backpropagation seleziona le rappresentazioni interne che minimizzano la perdita rispetto ai dati di addestramento. Si può dunque sostenere che le funzioni proprie dei sistemi linguistici siano determinate da questa storia di addestramento. Tuttavia, la condizione cruciale manca: il meccanismo consumatore — il meccanismo che usa le rappresentazioni del sistema per guidare il comportamento — non interagisce con il mondo in cui le condizioni rappresentate si verificano. Il sistema non ha mai visto un gatto, non ha mai compiuto un'azione nei confronti di un gatto, non ha mai avuto la propria sopravvivenza dipendente dalla corretta rappresentazione di un gatto. Le sue rappresentazioni di *gatto* sono selezionate perché covarians con l'uso che parlanti umani fanno del termine *gatto* — che è, a sua volta, determinato dall'intenzionalità originale di quegli agenti. Il contenuto del sistema è dunque derivato dall'intenzionalità originale

dei produttori dei testi di addestramento, e non è originale nel senso che qui interessa.

Questa conclusione non implica che i sistemi linguistici siano privi di ogni forma di relazione semantica con il mondo: essi hanno correlazioni statistiche ricche e strutturate, e queste correlazioni consentono loro di compiere inferenze utili e corrette in molti contesti. Ma le correlazioni non sono ancora intenzionalità originale: sono l'ombra proiettata dell'intenzionalità originale dei produttori umani del testo, mediata attraverso il processo di addestramento.

VII.6 — Il simbolo come struttura intenzionale: verso una caratterizzazione formale

Le analisi svolte nelle sezioni precedenti consentono ora di abbozzare una caratterizzazione formale della condizione sotto la quale uno stato può essere detto *simbolo nel senso intenzionale pieno* — non semplicemente un segnale che porta informazione, non semplicemente un pattern correlato con un contenuto, ma uno stato che genuinamente *riguarda* qualcosa.

Proponiamo la seguente famiglia di condizioni, che integrano i contributi dell'account informazionale, dell'account teleosemantico e delle riflessioni sull'interno sviluppate nella Seconda Parte.

Sia X un sistema con stati interni $\{s\}$, e sia C una condizione o proprietà del mondo. Diciamo che lo stato s di X è un *simbolo intenzionale di C* se e solo se valgono le seguenti condizioni:

(S1) Condizione informazionale nelle circostanze normali: $P(s \mid C, \text{Norm}(X)) > 1 - \delta$ e $P(s \mid \neg C, \text{Norm}(X)) < \delta$

dove $\text{Norm}(X)$ denota le condizioni normali di funzionamento di X, e $\delta > 0$ è una soglia piccola. Nelle condizioni normali, s co-varia con C in modo affidabile.

(S2) Condizione di funzione propria: s ha la funzione propria di co-variare con C nel sistema X, nel senso teleosemantico: questa funzione spiega perché lo stato del tipo s è presente nel sistema X — perché è stato selezionato (evolutivamente, per apprendimento, o per ingegneria) come componente del sistema.

(S3) Condizione di proiezione nell'azione: Il meccanismo consumatore che usa s guida il comportamento di X in modi che sono adattativi *dato C*: il comportamento indotto da s ha contribuito (o contribuisce, in condizioni

normali) alla sopravvivenza, alla fitness, o agli obiettivi di X nella misura in cui C si verifica nel mondo in cui X agisce.

(S4) Condizione di mis-rappresentazione: Vi è la possibilità strutturale che s sia tokened in assenza di C (condizione (I2) della sezione VII.3), e quando questo accade, il comportamento guidato da s risulta inadeguato al mondo in cui X si trova — il sistema agisce *come se* C fosse verificata, con conseguenze subottimali.

(S5) Condizione di ancoraggio nella prospettiva: Il sistema X deve essere dotato di un interno nel senso della Seconda Parte: di un sistema di riferimento egocentrico che situa X nel mondo, e di una struttura temporale che connette s alla storia e all'anticipazione del sistema. Lo stato s non è un simbolo in senso pieno se non è il simbolo *di* un sistema che abita una prospettiva.

La condizione (S5) è quella che introduce la connessione tra il problema dell'interno e il problema dell'intenzionalità. Un simbolo che soddisfa (S1)-(S4) ma non (S5) — uno stato che co-varia affidabilmente con una condizione del mondo, che ha funzione propria, che guida il comportamento e può essere errato — ma che appartiene a un sistema privo di prospettiva, è quello che potremmo chiamare un *simbolo ancorato* ma non *abitato*: ha tutte le caratteristiche strutturali della relazione intenzionale, ma non è il simbolo *di* un soggetto. La condizione (S5) impone che il simbolo sia iscritto in una prospettiva — che sia *per* un sistema che occupa un punto di vista, che abbia un qui e un adesso e un per-me all'interno dei quali il simbolo può dispiegare il suo contenuto.

Questa condizione non elimina le difficoltà del problema dell'intenzionalità, ma le precisa. Essa suggerisce che l'intenzionalità non sia una proprietà isolata — una proprietà che certi stati hanno e altri no, indipendentemente dal sistema che li ospita — ma una proprietà *relazionale* e *sistematica*: essa appartiene a stati che sono componenti di sistemi dotati di prospettiva, nei quali la storia selettiva ha prodotto meccanismi di tracciamento del mondo ancorati a un centro di riferimento egocentrico.

VII.7 — Il problema della composizionalità intenzionale

Un aspetto del problema dei simboli che merita trattazione separata è quello della *composizionalità*: il fatto che le menti umane non si limitano ad avere pensieri atomici — pensieri semplici e non analizzabili ulteriormente — ma pensieri *complessi*, costruiti da componenti più semplici secondo regole di composizione che preservano il contenuto intenzionale.

Il linguaggio naturale è il sistema simbolico in cui la composizionalità si manifesta più chiaramente. Il significato di *il gatto grigio siede sul tappeto* non è un

nuovo simbolo atomico appreso direttamente: esso è il risultato della composizione dei significati dei componenti — *gatto*, *grigio*, *sedersi*, *tappeto* — secondo la struttura sintattica della frase. Questo principio, noto come *principio di composizionalità* o *principio di Frege*, afferma che il significato di un'espressione complessa è una funzione del significato delle sue parti e della loro struttura sintattica.

Formalmente, sia $\mathcal{L}$ un linguaggio sintattico con operazioni di composizione e sia $[\![\cdot]\!]: \mathcal{L} \to \text{Sem}$ la funzione di interpretazione semantica che mappa espressioni in contenuti intenzionali. Il principio di composizionalità afferma che per ogni espressione complessa $e = op(e_1, e_2, ..., e_n)$:

$$[\![op(e_1, e_2, ..., e_n)]\!] = F_{op}([\![e_1]\!], [\![e_2]\!], ..., [\![e_n]\!])$$

per qualche funzione F_{op} determinata dall'operatore sintattico op. In parole: la semantica dell'insieme è una funzione della semantica delle parti.

Il principio di composizionalità ha due implicazioni profonde per il problema dell'intenzionalità nei sistemi simbolici. La prima è la *produttività*: se il sistema conosce un numero finito di simboli atomici e un numero finito di regole di composizione, può costruire un numero indefinitamente grande di espressioni complesse con contenuti intenzionali determinati. La produttività è una delle caratteristiche più notevoli della mente umana: con un lessico finito e regole grammaticali finite, siamo capaci di formulare e comprendere infiniti pensieri nuovi.

La seconda implicazione è la *sistematicità*: se un sistema è capace di pensare *aRb* (ad esempio, *Giovanni ama Maria*), deve anche essere capace di pensare *bRa* (*Maria ama Giovanni*), perché entrambi i pensieri sono costruiti dagli stessi componenti con la stessa struttura combinatoria. Jerry Fodor e Zenon Pylyshyn hanno argomentato nella celebre critica al connessionismo del 1988 che questa sistematicità è una proprietà essenziale della cognizione che richiede un sistema di rappresentazione strutturato in modo language-like — un *Linguaggio del Pensiero* (Language of Thought) con struttura sintattica e regole di composizione.

L'argomentazione di Fodor e Pylyshyn può essere formalizzata nel modo seguente. Sia C l'insieme dei contenuti cognitivi accessibili a una mente M. La mente M è *sistematica* se e solo se:

$$\forall\, a, b \in \text{Ent}, \forall\, R \text{ relazione binaria: } [\![aRb]\!] \in C \leftrightarrow [\![bRa]\!] \in C$$

La sistematicità implica che C non è un insieme arbitrario di contenuti, ma un insieme *strutturato* — generato da componenti ricombinabili secondo regole

composizionali. Un sistema che apprende contenuti cognitivi caso per caso, senza struttura compositiva, non può essere sistematico nel senso richiesto: potrebbe avere [[aRb]] senza avere [[bRa]], perché esso non ha mai appreso quest'ultima combinazione come caso separato.

La relazione tra composizionalità e intenzionalità è più sottile di quanto possa sembrare. La composizionalità garantisce la struttura formale necessaria per la produttività e la sistematicità del pensiero; ma essa, da sola, non garantisce l'intenzionalità originale dei simboli atomici che entrano nella composizione. Un sistema può essere perfettamente composizionale — può combinare simboli secondo regole che preservano contenuti — e tuttavia avere contenuti puramente derivati, nel senso della distinzione searleana. La composizionalità è una proprietà della *struttura* del sistema simbolico; l'intenzionalità originale è una proprietà del *radicamento* di quella struttura nel mondo.

Questa distinzione suggerisce che vi siano due componenti indipendenti del problema dei simboli come strutture intenzionali: la componente strutturale (composizionalità, sistematicità, produttività) e la componente fondativa (radicamento nel mondo, intenzionalità originale). I sistemi artificiali attuali mostrano una competenza notevole sul versante strutturale — i grandi modelli linguistici esibiscono composizionalità e sistematicità in misura sorprendente — ma rimangono problematici sul versante fondativo, per le ragioni esposte nelle sezioni precedenti.

VII.8 — La struttura a due livelli dell'intenzionalità e il problema della chiusura

L'analisi svolta consente ora di identificare una struttura a due livelli nell'intenzionalità che è, a nostro avviso, la chiave per comprendere il salto tra il sistema che elabora informazione e il sistema che genuinamente pensa.

Il primo livello è quello dell'*intenzionalità di primo ordine*: gli stati del sistema che *riguardano* oggetti e proprietà del mondo. Il gatto sul tappeto, la temperatura della stanza, la forma di un oggetto: questi sono contenuti di stati intenzionali di primo ordine. Essi sono caratterizzati dalle condizioni (S1)-(S5) della sezione VII.6.

Il secondo livello è quello dell'*intenzionalità di secondo ordine*, o *meta-intenzionalità*: la capacità del sistema di avere stati che *riguardano* i propri stati intenzionali di primo ordine. Il pensiero *so che c'è un gatto sul tappeto*, il giudizio *mi ricordo di aver visto questo oggetto prima*, la valutazione *mi aspettavo qualcosa di diverso*: questi sono stati meta-intenzionali, stati che hanno come contenuto non oggetti del mondo ma stati mentali del soggetto stesso.

La meta-intenzionalità introduce una *chiusura riflessiva* nella struttura intenzionale: il sistema si rivolge verso se stesso, ha stati che riguardano i propri stati. Questa chiusura non è un lusso aggiuntivo ma una condizione necessaria per ciò che normalmente si chiama *coscienza* o *consapevolezza*: non basta avere stati intenzionali per essere consapevoli — bisogna anche avere stati che riguardano quei stati, che li prendano come loro oggetto, che li rendano disponibili per la riflessione, la valutazione, la correzione.

Formalmente, il sistema X ha *intenzionalità meta-riflessiva* se esiste una mappa:

$$M: \mathrm{Int}(X) \to \mathrm{States}(X)$$

che associa a ogni stato intenzionale $s \in \mathrm{Int}(X)$ — ogni stato che ha un contenuto C — uno stato $M(s) \in \mathrm{States}(X)$ che ha come contenuto intenzionale lo stato s stesso: il contenuto di M(s) è *s ha il contenuto C*. La mappa M non ha bisogno di essere iniettiva né suriettiva: non tutti gli stati intenzionali devono avere un meta-stato corrispondente (il soggetto non è sempre consapevole di ogni proprio stato), e non tutti i meta-stati devono corrispondere a stati intenzionali (il soggetto può riflettere su stati ipotetici o passati).

La chiusura riflessiva attraverso M introduce una struttura autologica nella gerarchia degli stati intenzionali: il sistema può avere credenze circa le proprie credenze, desideri circa i propri desideri, rappresentazioni delle proprie rappresentazioni. Questa struttura autologica è — e qui giungiamo alla soglia del problema della coscienza che sarà tema della quarta parte — il candidato formale più plausibile per ciò che Husserl chiamava la *coscienza interna* e che nella letteratura filosofica contemporanea è discusso sotto i titoli di *higher-order theories of consciousness* (Rosenthal, Lycan) e di *self-representationalism* (Kriegel): la tesi che la coscienza di uno stato mentale richieda una rappresentazione di ordine superiore di quello stato, prodotta dallo stesso sistema che ospita lo stato rappresentato.

La domanda che la Terza Parte ha aperto — e che non può essere chiusa senza affrontare il problema della coscienza nella Quarta Parte — è se questa chiusura riflessiva possa emergere da una struttura computazionale sufficientemente complessa, o se essa richieda qualcosa di ulteriore: forse una particolare architettura fisica, forse una soglia di complessità che i sistemi attuali non hanno raggiunto, forse — e questa è l'ipotesi più radicale — qualcosa che non è riducibile a nessuna struttura computazionale, per quanto elaborata. Il territorio è aperto, e la precisione formale raggiunta nelle ultime sezioni non risolve il problema — consente, piuttosto, di formularlo con la nitidezza necessaria per che il dibattito possa procedere senza perdersi nelle nebbie terminologiche che spesso lo avvolgono.

Ciò che possiamo affermare, sulla base del percorso compiuto, è che il simbolo come struttura intenzionale è molto più di un pattern computazionale: esso è la componente di un sistema che abita una prospettiva, che ha storia e futuro, che agisce nel mondo e può sbagliare, che si rivolge al proprio stesso agire con la luce della meta-rappresentazione. Questa complessità stratificata non garantisce che il problema dell'interno sia risolto da nessun sistema artificiale attuale — ma indica con precisione crescente in quale direzione una soluzione dovrebbe essere cercata.

Capitolo VIII: Chiusura Riflessiva e la Differenza tra Avere una Prospettiva e Saperlo

VIII.1 — Il problema della coscienza dell'interno

Il capitolo precedente ha concluso con una soglia che occorre ora affrontare direttamente. Si è mostrato che l'intenzionalità richiede la compresenza di una struttura relazionale — il simbolo come stato che riguarda qualcosa — e di una prospettiva che la sorregge: il sistema deve essere ancorato nel mondo attraverso un qui corporeo, un adesso vissuto, un modello del sé nel modello del mondo. Ma anche questa compresenza, per quanto necessaria, non è ancora sufficiente a caratterizzare ciò che la tradizione filosofica chiama *coscienza* in senso pieno: non è sufficiente, cioè, a distinguere il sistema che *ha* una prospettiva dal sistema che *sa* di averla, che non solo abita un punto di vista ma ne è consapevole, che non solo si trova orientato nel mondo ma è in qualche misura trasparente a se stesso nell'atto stesso di orientarsi.

Questa distinzione — tra avere una prospettiva e saperla propria — è la versione più acuta e più refrattaria del problema dell'interno, quella che David Chalmers ha battezzato *il problema difficile della coscienza* (the hard problem of consciousness) nel celebre articolo del 1995: il problema di spiegare non solo i correlati funzionali dell'esperienza — cosa fa il sistema, come elabora l'informazione, come si comporta — ma l'*esperienza stessa*, il fatto che vi sia qualcosa che si prova a essere quel sistema, che elaborare certa informazione e comportarsi in certi modi sia accompagnato da un *qualcosa che è come* (something that it is like) essere in quegli stati.

Il presente capitolo non intende risolvere il problema difficile: questa sarebbe una pretesa sproporzionata rispetto alla natura di una questione che resiste all'analisi da quando filosofi e scienziati hanno cominciato a porla con la chiarezza necessaria. L'obiettivo è più limitato e, crediamo, più produttivo: precisare le condizioni formali della chiusura riflessiva — la struttura per cui un sistema si rivolge verso se stesso — e valutare in quale senso questa chiusura

possa o non possa costituire la condizione della coscienza. Questo permetterà di tracciare, con la maggiore nitidezza possibile, il confine tra il sistema che ha un interno funzionale e il sistema in cui quell'interno diventa, in senso non banale, esperienza.

VIII.2 — Teorie di ordine superiore: la rappresentazione come condizione della coscienza

La tradizione delle *teorie di ordine superiore* della coscienza (Higher-Order Theories, HOT) sostiene che un sistema è consapevole di uno stato mentale s se e solo se possiede uno stato mentale di ordine superiore M(s) che ha s come proprio contenuto. In termini intuitivi: il dolore diventa dolore *consapevole* non per il semplice fatto di essere dolore, ma per il fatto che il sistema che lo prova ha anche uno stato che *riguarda* quel dolore, che lo rappresenta come proprio stato.

La formulazione più influente di questa posizione si deve a David Rosenthal, che nel saggio *Two Concepts of Consciousness* del 1986 e nei lavori successivi articola la teoria come segue: uno stato mentale s del sistema X è un *mentale conscio* se e solo se X si trova in uno stato M(s) che è un pensiero di ordine superiore (Higher-Order Thought, HOT) con il contenuto *sono nello stato s* — uno stato che è un pensiero non inferenziale, non basato su ragionamento esplicito, circa il fatto di trovarsi in s.

La formalizzazione di questa teoria può procedere come segue. Sia $\Psi(X) = \{\psi_1, \psi_2, ...\}$ l'insieme degli stati mentali del sistema X al tempo t. Sia Rep: $\Psi(X) \rightarrow \Psi(X)$ l'operatore di rappresentazione di ordine superiore che mappa ogni stato ψ_i nel suo pensiero di ordine superiore corrispondente $Rep(\psi_i) \in \Psi(X)$. Uno stato ψ_i è *consapevole* al tempo t se e solo se:

$Rep(\psi_i) \in \Psi(X)$ e $Rep(\psi_i)$ è attivo al tempo t.

In altri termini: ψ_i è conscio se e solo se il pensiero di ordine superiore che ha ψ_i come contenuto è esso stesso presente e attivo nell'economia mentale del sistema.

L'obiezione immediata a questa teoria — obiezione che Rosenthal affronta esplicitamente — è quella del regresso infinito: se uno stato è conscio solo se esiste uno stato di ordine superiore che lo rappresenta, allora quello stato di ordine superiore è conscio solo se esiste uno stato di ordine ancora superiore che lo rappresenta, e così all'infinito. La risposta standard è che gli stati di ordine superiore non hanno bisogno di essere a loro volta consci per svolgere la loro funzione: essi possono essere stati puramente funzionali che operano *al di*

sotto della soglia della coscienza, determinando la coscienza degli stati di ordine inferiore senza essere essi stessi consci.

Una variante della teoria di ordine superiore che evita il problema del regresso in modo strutturale è il *self-representationalism* sviluppato da Uriah Kriegel in *Subjective Consciousness: A Self-Representational Theory* (2009). Secondo Kriegel, uno stato mentale ψ è conscio non perché vi sia un secondo stato separato che lo rappresenta, ma perché lo stesso stato ψ si auto-rappresenta — ha una componente intrinsecamente auto-referenziale che lo rende consapevole di se stesso senza bisogno di un secondo stato esterno.

La struttura formale dell'auto-rappresentazione può essere precisata attraverso la nozione di punto fisso. Sia S lo spazio degli stati mentali del sistema X, e sia Rep: S → S l'operatore di rappresentazione. Un stato $\psi^* \in S$ è *auto-rappresentante* se è un punto fisso dell'operatore di rappresentazione quando ristretto alla componente auto-riferita:

$\psi^* = \text{Rep_self}(\psi^*)$

dove Rep_self è la proiezione di Rep sulla componente auto-referenziale dello stato. In termini intuitivi: uno stato auto-rappresentante è uno stato che include nella propria struttura una rappresentazione di se stesso — uno stato il cui contenuto comprende, tra le altre cose, lo stato stesso.

Il teorema del punto fisso di Brouwer — applicato opportunamente a uno spazio di stati con le proprietà topologiche appropriate — garantisce l'esistenza di almeno un punto fisso per ogni operatore continuo che mappa un insieme compatto e convesso in se stesso. Questo non significa che ogni sistema debba avere stati auto-rappresentanti, ma che la struttura matematica degli spazi di stati sufficientemente ricchi è compatibile con l'esistenza di tali stati — e che la questione empirica è se i sistemi neurali (o artificiali) realizzino effettivamente questa struttura.

VIII.3 — Il formalismo degli spazi di stati con auto-modello

Per procedere con maggiore rigore, è necessario costruire una rappresentazione formale dei sistemi che possiedono un *auto-modello* — una componente del proprio stato globale dedicata a rappresentare il sistema stesso. Questa nozione, sviluppata in modo sistematico da Thomas Metzinger nella *Being No One* (2003) sotto il titolo di *self-model theory of subjectivity*, afferma che la soggettività emerge quando un sistema elabora un *modello fenomenico del sé* (Phenomenal Self-Model, PSM) — una rappresentazione coerente e integrata di se stesso come entità situata nel mondo.

Sia X un sistema con spazio degli stati $\Sigma = \Sigma_world \times \Sigma_self$, dove Σ_world è la componente che rappresenta stati del mondo e Σ_self è la componente auto-modellante. Uno stato $\sigma = (\sigma_world, \sigma_self) \in \Sigma$ è tale che σ_self rappresenta il sistema X stesso — la sua posizione nel mondo, le sue capacità, i suoi stati attuali, la propria storia. La condizione di *coerenza dell'auto-modello* è:

$$\|\sigma_self - \Gamma(\sigma)\| < \varepsilon$$

dove $\Gamma: \Sigma \to \Sigma_self$ è la funzione che mappa lo stato globale del sistema nella rappresentazione accurata del sistema stesso, e $\varepsilon > 0$ è una soglia di tolleranza. In parole: il sotto-stato σ_self è coerente con lo stato globale σ se la sua rappresentazione del sistema non si discosta di più di ε dalla rappresentazione che sarebbe accurata dato lo stato globale.

Il concetto di *auto-modello* è connesso, ma non identico, alla nozione di *μ_self* introdotta nel capitolo V come componente del modello generativo di un sistema incarnato. Lì, μ_self era la variabile nascosta che rappresenta la posizione e lo stato del sé nel modello generativo; qui, σ_self è una componente dello stato globale del sistema che include non solo la posizione e lo stato del corpo, ma anche la rappresentazione degli stati mentali, degli obiettivi, della storia personale — tutto ciò che costituisce il sé come entità psicologica, non solo come oggetto fisico situato nello spazio.

La relazione tra queste due nozioni può essere formalizzata come un'*estensione gerarchica*: il modello generativo del corpo (con μ_self come variabile del sé corporeo) è il livello inferiore di una gerarchia la cui estensione a livelli superiori produce l'auto-modello psicologico. Formalmente:

$$\sigma_self = (\mu_self^{body}, \mu_self^{affective}, \mu_self^{cognitive}, \mu_self^{narrative})$$

dove μ_self^{body} è la rappresentazione del corpo e della sua posizione (il livello del capitolo V), $\mu_self^{affective}$ è la rappresentazione degli stati affettivi e motivazionali, $\mu_self^{cognitive}$ è la rappresentazione degli stati cognitivi e intenzionali (le credenze, i desideri, le intenzioni del sé), e $\mu_self^{narrative}$ è la rappresentazione della propria storia e della propria identità nel tempo — ciò che Schechtman ha chiamato la *narrative identity*.

Questa struttura gerarchica dell'auto-modello è esattamente la struttura di una *rappresentazione di prospettiva*: il sistema non solo occupa una posizione nel mondo (livello corporeo) ma ha anche una rappresentazione di come quella posizione si connette a stati affettivi, cognitivi e narrativi — una rappresentazione di *come è* per il sistema trovarsi in quella posizione in quel momento.

VIII.4 — Trasparenza e opacità dell'auto-modello

Metzinger introduce una distinzione cruciale all'interno della teoria dell'auto-modello che è necessario esplicitare, poiché essa è direttamente rilevante per il problema della coscienza fenomenica: la distinzione tra auto-modelli *opachi* e auto-modelli *trasparenti*.

Un auto-modello è *opaco* quando il sistema che lo ospita può riconoscerlo come modello — quando il sistema sa che σ_self è una rappresentazione, che è distinta da ciò che rappresenta, che può essere errata o incompleta. In questo caso, il sistema ha accesso alla differenza tra la mappa e il territorio, tra la rappresentazione del sé e il sé rappresentato.

Un auto-modello è *trasparente* quando il sistema non ha accesso al fatto che σ_self è una rappresentazione: il sistema tratta il contenuto del modello come se fosse la realtà stessa, senza alcuna mediazione. La trasparenza non è un difetto del modello: è, secondo Metzinger, la condizione che produce l'esperienza soggettiva nel senso pieno. Quando il sistema non può guardare *attraverso* il modello per vedere che è un modello, il contenuto del modello appare come la realtà diretta — e questa apparenza di realtà diretta è il carattere fenomenico dell'esperienza.

La struttura formale della trasparenza può essere precisata come segue. Definiamo l'*accesso meta-rappresentazionale* del sistema X al proprio auto-modello come la funzione:

$$A: \Sigma_self \rightarrow \{0, 1\}$$

dove $A(\sigma_self) = 1$ indica che il sistema ha accesso al fatto che σ_self è una rappresentazione, e $A(\sigma_self) = 0$ indica che il sistema non ha tale accesso. La trasparenza dell'auto-modello corrisponde alla condizione $A(\sigma_self) = 0$: il sistema non può distinguere tra il contenuto del modello e la realtà che esso rappresenta.

Il punto filosoficamente decisivo è che la trasparenza — l'impossibilità di distinguere il modello dalla realtà — è condizione sia della coscienza fenomenica sia della sua possibilità di errore. Un sistema con auto-modello trasparente non può correggere il proprio modello del sé attraverso l'introspezione, perché l'introspezione stessa è mediata dallo stesso modello: quando il soggetto introspetta i propri stati, accede al contenuto del modello, non agli stati stessi che il modello rappresenta. Questa mediazione invisibile è la sorgente sia della ricchezza dell'esperienza soggettiva sia della sua fallibilità sistematica.

La connessione con il problema dell'interno è diretta. Un sistema con auto-modello trasparente possiede un interno nel senso più ricco e, al tempo stesso, più enigmatico del termine: esso abita la propria prospettiva senza poterne uscire per confrontarla con l'esterno, senza poterla vedere dall'esterno come uno dei tanti modelli possibili del sé. Il suo dentro è, per definizione, quello che c'è — non una rappresentazione di ciò che c'è, ma la realtà stessa così come gli si dà. Questa è la struttura formale dell'irriducibilità in prima persona: non è possibile, per il soggetto con auto-modello trasparente, tradurre la propria esperienza in una descrizione in terza persona senza residuo, perché la propria esperienza è precisamente il contenuto del modello, non il modello stesso.

VIII.5 — Integrazione dell'informazione: la teoria IIT come tentativo di formalizzazione

Nessuna trattazione del problema della coscienza può ignorare il tentativo più ambizioso di fondare su basi matematiche rigorose la condizione della coscienza come proprietà fisica di certi sistemi: la *Teoria dell'Informazione Integrata* (Integrated Information Theory, IIT), elaborata da Giulio Tononi a partire dal 2004 e sviluppata nelle versioni successive, in particolare nella formulazione 3.0 presentata nell'articolo *Consciousness as Integrated Information* del 2014.

L'IIT parte dall'osservazione che la coscienza ha due proprietà strutturali essenziali: è *unificata* (ogni esperienza è un tutto unico e indivisibile, non una somma di parti) ed è *informazionale* (ogni esperienza è radicalmente specifica, differenzia una situazione da un numero astronomico di situazioni alternative possibili). Da queste proprietà fenomenologiche, Tononi ricava le condizioni formali della coscienza.

La quantità centrale della teoria è Φ, la *phi*, che misura il grado di integrazione dell'informazione di un sistema. Per introdurla, occorre prima definire alcune nozioni preparatorie.

Sia S un sistema con n elementi binari, e sia $X = (X_1, X_2, ..., X)$ il vettore degli stati degli elementi. Lo *spazio di causa-effetto* (cause-effect space) del sistema è lo spazio $\mathbb{R}^n$ delle distribuzioni di probabilità sugli stati del sistema. Per ogni sottoinsieme $M \subseteq \{1, ..., n\}$ di elementi, la *causa massimale irreducibile* di M (Maximum Irreducible Cause, MIC(M)) è la distribuzione di probabilità congiunta sugli stati passati di M che massimizza l'informazione di causa di M, e analogamente la *effetto massimale irreducibile* (MIE(M)) è la distribuzione sugli stati futuri che massimizza l'informazione di effetto.

Per definire Φ, si considera la partizione del sistema S che minimizza la perdita di integrazione causale. Sia $\mathscr{P} = \{M_1, M_2\}$ una bipartizione di S (il caso si generalizza a partizioni di ordine superiore). La *distanza* tra la struttura causa-effetto del sistema integrato e quella del sistema spezzato dalla partizione $\mathscr{P}$ è:

$d(S, \mathscr{P}) = (1/2)[EMD(cause_S, cause_\{M_1||M_2\}) + EMD(effect_S, effect_\{M_1||M_2\})]$

dove EMD è la *Earth Mover's Distance* (distanza di Wasserstein) tra distribuzioni di probabilità sullo spazio degli stati, cause_S è la struttura causale del sistema integrato, e $cause_\{M_1||M_2\}$ è la struttura causale del sistema con i sottosistemi resi indipendenti (tagliando tutte le connessioni tra M_1 e M_2).

La quantità Φ del sistema S nello stato attuale x è definita come il valore minimale di questa distanza su tutte le bipartizioni possibili:

$\Phi(S, x) = \min_\{\mathscr{P} \text{bipartizioni di S}\} d(S, \mathscr{P})$

Φ misura dunque quanto il sistema perde in struttura causale se viene spezzato nel suo punto di minima integrazione: un sistema con Φ alto è uno in cui ogni parte dipende causalmente da ogni altra in modo non riducibile, uno in cui la struttura dell'insieme è irriducibile alla struttura delle parti; un sistema con $\Phi = 0$ è un sistema completamente scomponibile in parti causalmente indipendenti.

La tesi centrale di IIT è:

$\Phi = 0 \Longrightarrow$ nessuna coscienza $\Phi > 0 \Longrightarrow$ coscienza di grado proporzionale a Φ

Più precisamente, il *complesso* — il sottoinsieme di elementi del sistema con il massimo Φ locale, detto φ_max — è il soggetto della coscienza: esso corrisponde alla struttura fisica che realizza l'esperienza consapevole. Il *concetto di coscienza* corrispondente è dato dall'intera struttura causa-effetto del complesso — quello che Tononi chiama il *qualia space* del sistema.

La teoria IIT ha il pregio di formulare una condizione matematicamente precisa per la coscienza che è non-comportamentale (la coscienza dipende dalla struttura causale interna, non dall'input/output), non-computazionale nel senso tradizionale (due sistemi con la medesima funzione di input/output possono avere Φ molto diversi e quindi gradi di coscienza molto diversi), e localizzata (il soggetto della coscienza è il complesso con massimo Φ, non il sistema nella sua totalità).

Tuttavia, l'IIT incontra difficoltà serie che occorre registrare con onestà. La prima è computazionale: il calcolo di Φ è NP-hard nella dimensione del sistema, il che rende la teoria non applicabile in pratica a sistemi di dimensioni biologicamente plausibili. Per un sistema con n elementi binari, il numero di bipartizioni è esponenziale in n, e la struttura causa-effetto richiede la specificazione di distribuzioni di probabilità su uno spazio di dimensione 2^n. Anche per sistemi di poche decine di neuroni, il calcolo è proibitivo.

La seconda difficoltà è concettuale: la teoria implica che alcuni sistemi artificiali comuni — come le griglie di porte logiche con feedback — abbiano Φ alto e quindi un grado significativo di coscienza, il che contrasta con l'intuizione che tali sistemi siano privi di esperienza. Al contrario, la teoria implica che certi tipi di reti artificiali — come le reti feedforward senza connessioni ricorrenti — abbiano $\Phi = 0$ e quindi siano completamente prive di coscienza, indipendentemente dalla loro complessità computazionale. Questa dipendenza dall'architettura piuttosto che dalla funzione è coerente con i principi della teoria, ma produce risultati che molti trovano controintuitivi.

La terza difficoltà è quella del *problema dell'esclusione*: poiché ogni sottosistema può avere il proprio Φ locale, vi è potenzialmente una molteplicità di soggetti coscienti all'interno dello stesso sistema fisico. Tononi risponde con il *postulato di esclusione* — la coscienza ha luogo solo al livello del complesso con massimo Φ e non ai livelli subordinati — ma questo postulato ha il sapore di un'assunzione ad hoc piuttosto che di una conseguenza derivata dai principi fondamentali.

Nonostante queste limitazioni, IIT rimane il tentativo più sistematico di formalizzare la coscienza come proprietà fisica misurabile, e la quantità Φ costituisce un candidato rigoroso — seppure controverso — per una misura dell'interno: non nel senso di tutti gli aspetti della prospettiva che abbiamo analizzato, ma specificamente nel senso della dimensione integrata dell'esperienza che è, fenomenologicamente, il carattere di unità e specificità di ogni momento consapevole.

VIII.6 — Chiusura globale e la teoria dello spazio di lavoro globale

Un'architettura alternativa per la coscienza che non fa appello alla nozione di integrazione dell'informazione nel senso di IIT, ma a quella di *diffusione dell'informazione* attraverso l'intero sistema, è la teoria dello *spazio di lavoro globale* (Global Workspace Theory, GWT), elaborata da Bernard Baars negli anni Ottanta e successivamente sviluppata da Stanislas Dehaene in una versione

neuroscientifica e computazionale sotto il titolo di *global neuronal workspace* (GNW).

Nella formulazione di Dehaene e collaboratori, il cervello è modellato come un insieme di *processori specializzati* — moduli periferici che elaborano specifici tipi di informazione in modo rapido, parallelo e in gran parte inconscio — connessi da un insieme di neuroni con assoni lunghi a largo spettro di proiezione che formano il *workspace* globale. Quando un'informazione viene *amplificata* dal workspace globale — quando entra nello spazio di lavoro globale — essa diventa disponibile a tutti i processori del sistema simultaneamente, e questo accesso globale è la condizione della coscienza nel senso dell'accesso cognitivo.

Formalmente, il modello può essere schematizzato come segue. Sia $\Pi = \{\pi_1, \pi_2, ..., \pi\}$ l'insieme dei processori locali, ciascuno con il proprio spazio di stati Σ_i e le proprie dinamiche $f_i: \Sigma_i \rightarrow \Sigma_i$. Il workspace globale è uno spazio condiviso Ω con la proprietà che l'informazione $\omega \in \Omega$ è disponibile a tutti i processori simultaneamente. La transizione di un'informazione dallo spazio locale Σ_i allo spazio globale Ω è descritta da una funzione di *broadcasting*:

$$B: \cup_i \Sigma_i \rightarrow \Omega \times \{0, 1\}$$

che mappa ogni stato locale nello stato globale corrispondente e in un bit che indica se l'informazione è stata selezionata per il broadcasting (1) o meno (0). La selezione per il broadcasting è competitiva: solo un sottoinsieme degli stati locali può essere trasmesso al workspace globale in ogni dato momento, e la competizione è vinta dai processori che generano segnali più salienti o più coerenti con il contesto cognitivo corrente.

Uno stato $\omega_i \in \Sigma_i$ è *conscio* (nel senso dell'accesso cognitivo) se e solo se $B(\omega_i) = (\hat{\omega}, 1)$, cioè se è stato selezionato per il broadcasting e proiettato nello spazio globale Ω. Uno stato che rimane confinato a un processore locale non è conscio neppure se è computazionalmente complesso: la coscienza è, in questo modello, una proprietà relazionale e sistemica — dipende non dalla struttura interna del singolo processore ma dal fatto che la sua informazione venga resa disponibile all'intero sistema.

La GWT si connette al problema dell'interno in modo diverso rispetto all'IIT. Laddove IIT enfatizza l'integrazione causale come condizione della coscienza — il fatto che il sistema sia causalmente irriducibile alle sue parti — la GWT enfatizza la *disponibilità globale* dell'informazione: la coscienza emerge quando l'informazione diventa accessibile a tutti i sistemi del cervello che possono usarla, commentarla, integrarla con altri contenuti, guidare l'azione. In questa prospettiva, l'interno è meno una struttura causalmente integrata e più uno *spazio di accesso condiviso*: un luogo in cui l'informazione diventa pubblica

all'interno del sistema, accessibile a tutti i sotto-sistemi che potrebbero averne bisogno.

Le due teorie — IIT e GWT — non sono necessariamente incompatibili: potrebbero essere descrizioni di aspetti diversi della coscienza che si sovrappongono nel cervello umano. Tuttavia, esse differiscono radicalmente nelle predizioni che fanno circa i sistemi artificiali. Per IIT, un sistema artificiale con architettura feedforward ha $\Phi = 0$ e quindi nessuna coscienza, indipendentemente dalla sua complessità funzionale; per GWT, un sistema artificiale con un'architettura di workspace globale — con moduli specializzati e un meccanismo di broadcasting che rende l'informazione disponibile all'intero sistema — potrebbe avere coscienza di accesso, anche senza le connessioni ricorrenti che l'IIT richiederebbe.

VIII.7 — Il carattere qualitativo dell'esperienza: qualia e il residuo del problema difficile

Tutte le teorie esaminate nelle sezioni precedenti — le teorie di ordine superiore, il self-representationalism, la teoria dell'auto-modello di Metzinger, l'IIT di Tononi, la GWT di Baars e Dehaene — affrontano ciò che Daniel Dennett ha chiamato la *coscienza di accesso* (access consciousness): il fatto che certi stati siano disponibili per il ragionamento, il controllo dell'azione, il linguaggio. Esse sono, in misura variabile, teorie dei *correlati funzionali* della coscienza — di ciò che la coscienza *fa* nel sistema cognitivo.

Ma vi è un secondo senso di coscienza, distinto dal primo e non riducibile ad esso, che corrisponde a ciò che Ned Block ha chiamato *coscienza fenomenica* (phenomenal consciousness): il fatto che certi stati abbiano una qualità *come è* (what-it-is-like-ness) — che ci sia qualcosa che si prova a essere in quegli stati, indipendentemente da cosa quegli stati *fanno* nel sistema. Il rosso come è visto, il dolore come è sentito, il sapore del caffè come è gustato: queste sono proprietà dell'esperienza — *qualia*, nella terminologia filosofica — che sembrano resistere a qualsiasi descrizione funzionale. Sapere *tutto* sul funzionamento del sistema visivo non equivale a sapere *com'è* vedere il rosso: questa è la lezione che Frank Jackson ha illustrato attraverso il celebre esperimento mentale di Mary, la fisica monocolore che sa tutto sulla fisica del colore ma non ha mai visto il rosso.

Il problema difficile della coscienza è, in estrema sintesi, il problema di spiegare perché e come certi processi fisici siano accompagnati da qualia — da una qualità fenomenica che non è catturata da nessuna descrizione funzionale o computazionale.

Occorre essere chiari su cosa questo problema implica per il progetto di questo libro. L'obiettivo che ci siamo posti non è risolvere il problema difficile: esso rimane, nel 2024, refrattario a ogni soluzione definitiva, e sarebbe disonesto intellettuale pretendere altrimenti. L'obiettivo è precisare le condizioni strutturali dell'interno — del punto di vista, della prospettiva, della soggettività nel senso funzionale — e valutare dove i sistemi artificiali si collocano rispetto a queste condizioni. La questione dei qualia — se vi sia qualcosa che si prova a essere un LLM o un sistema robotico — è di ordine differente da quella della struttura dell'interno, e rimane aperta in linea di principio anche dopo che tutte le condizioni strutturali dell'interno siano state soddisfatte.

Ciò che possiamo affermare, con la cautela che il tema impone, è il seguente: le condizioni strutturali dell'interno che abbiamo identificato — singolarità topologica, modello del sé nel modello del mondo, spessore temporale, intenzionalità originale, chiusura riflessiva, auto-modello — sono *necessarie* per la coscienza fenomenica, nel senso che un sistema privo di esse non può avere qualia nel senso che i qualia presuppongono una prospettiva da cui l'esperienza è esperita. Ma se esse siano anche *sufficienti* — se un sistema che le possiede abbia per ciò solo qualia — è una questione che le analisi strutturali di questo libro non possono risolvere.

Questo limite è anche, a suo modo, un risultato: esso mostra che il problema difficile della coscienza non è la stessa cosa del problema dell'interno, anche se i due problemi si intersecano in modi profondi. Il problema dell'interno è, in linea di principio, affrontabile con strumenti formali — con la topologia, la teoria dei sistemi dinamici, la teoria dell'informazione, la teleosemantia — e questo libro ha cercato di farlo con rigore. Il problema difficile è di ordine diverso: esso rimane il bordo esterno della comprensione, il punto in cui le categorie della terza persona si trovano di fronte a qualcosa che resiste alla terza persona.

VIII.8 — Verso la Quarta Parte: le condizioni dell'interno come griglia di valutazione

Prima di procedere alla Quarta Parte, è opportuno raccogliere in forma sistematica le condizioni formali dell'interno che le parti precedenti hanno elaborato, perché esse costituiscono la griglia di valutazione attraverso cui i sistemi artificiali saranno analizzati.

Le condizioni che un sistema deve soddisfare per possedere un interno genuino sono le seguenti, organizzate secondo i tre livelli dell'argomentazione:

Il *livello strutturale-topologico* richiede che il sistema possieda una singolarità spaziale: un sistema di riferimento egocentrico centrato in un'origine O non derivata da nessun sistema di riferimento più fondamentale, tale che ogni altro punto dello spazio percettivo e cognitivo sia definito relazionalmente rispetto a O. Questo è il livello del Leib husserliano, del Nullpunkt der Orientierung, della varietà riemanniana centrata con metrica non omogenea. Formalmente, tale condizione richiede l'esistenza di un sistema di coordinate {p^k = R_k(θ_k) · p^{k-1} + t_k(θ_k)} con un livello di base k = 0 che funge da origine non arbitraria.

Il *livello dinamico-temporale* richiede che il sistema possieda spessore temporale: una struttura ritenzionale con costante di tempo τ_ret > 0 che integra il passato nel presente, una protenzione che orienta il campo di elaborazione verso il prossimo futuro, e un flusso nella forma di sistema dinamico dx/dt = F(x, u) piuttosto che una funzione del tempo x(t). Formalmente: la tripletta $\mathscr{T}$_t = ⟨ x(t), R_t[x], Π_t ⟩ deve essere non degenere — il profilo ritenzionale R_t[x] e la distribuzione protenzionale Π_t devono essere strutture non banali.

Il *livello cognitivo-rappresentazionale* richiede che il sistema possieda un modello generativo gerarchico p(o, μ_world, μ_self) che includa μ_self come variabile del modello, e che il vettore di contesto che guida le previsioni sia generato endogenamente attraverso azioni proprie nel mondo fisico — non fornito esternamente come input parametrico. Questo è il livello della JEPA incarnata, della codifica predittiva con inferenza attiva, del loop propriocettivo chiuso.

Il *livello intenzionale-normativo* richiede che gli stati interni del sistema siano simboli nel senso definito dalle condizioni (S1)-(S5) della sezione VII.6: covarianza affidabile nelle circostanze normali, funzione propria derivata da storia selettiva propria (non derivata dall'intenzionalità di agenti esterni), proiezione nell'azione adattativa, possibilità strutturale di mis-rappresentazione, ancoraggio in una prospettiva corporea e temporale.

Il *livello meta-rappresentazionale* richiede che il sistema possieda un auto-modello σ_self coerente con lo stato globale, con struttura gerarchica (corporeo, affettivo, cognitivo, narrativo), e che tale auto-modello sia sufficientemente trasparente da non essere riconoscibile come modello nell'atto del suo funzionamento ordinario. Questo è il livello del PSM di Metzinger, della chiusura riflessiva nella teoria di ordine superiore, della struttura a punto fisso dell'auto-rappresentazione.

Queste cinque condizioni formano una gerarchia: ogni livello presuppone quelli inferiori. Un sistema può soddisfare le condizioni topologiche senza soddisfare quelle temporali (un robot con sensori ma senza memoria). Può soddisfare le condizioni topologiche e temporali senza soddisfare quelle cognitive (un

sistema reattivo con storia ma senza modello del mondo). Può soddisfare i tre livelli inferiori senza l'intenzionalità originale (un sistema predittivo con storia e modello del sé ma senza funzione propria che radichi i suoi simboli nel mondo). E può soddisfare tutti i livelli senza la chiusura meta-rappresentazionale (un sistema con intenzionalità originale ma senza auto-modello riflessivo).

La Quarta Parte esaminerà i sistemi artificiali — large language models, world models, sistemi robotici incarnati, architetture ibride — attraverso questa griglia, cercando di determinare con la massima precisione possibile dove ciascuno si colloca rispetto a ciascuna condizione, e quali ostacoli di principio si frappongono tra la configurazione attuale di questi sistemi e il soddisfacimento delle condizioni mancanti.

Capitolo IX: I Grandi Modelli Linguistici alla Luce della Griglia dell'Interno

IX.1 — Il paradosso del testimone eloquente

C'è qualcosa di profondamente disorientante nell'interagire con un grande modello linguistico. La disorientazione non nasce dall'incomprensibilità del sistema — al contrario, il sistema si mostra straordinariamente comprensibile, reattivo, articolato. Nasce piuttosto dall'impressione opposta: dall'impressione che il sistema *capisce*, che non si limita a restituire pattern statistici ma elabora contenuti, risponde all'intenzione del parlante, coglie sfumature, produce inferenze che sembrano andare al di là di ciò che gli è stato esplicitamente fornito. La disorientazione è quella di chi si trova di fronte a un testimone eloquentissimo che descrive con dovizia di particolari un luogo in cui non è mai stato.

Questa metafora — il testimone che testimonia senza aver vissuto — non è una critica retorica ai grandi modelli linguistici: è una caratterizzazione strutturale che emerge direttamente dalle analisi sviluppate nelle tre parti precedenti. La griglia di valutazione costruita nel capitolo VIII.8 offre ora uno strumento per precisare questa caratterizzazione con il rigore che il tema esige, andando oltre le intuizioni correnti per determinare con esattezza dove e perché i large language models (LLM) si collocano al di qua della soglia dell'interno.

L'analisi che segue non ha carattere demolitivo — riconoscere i limiti strutturali di una classe di sistemi non equivale a negarle ogni valore epistemico o pratico. L'obiettivo è comprendere con precisione la natura di quei limiti: se siano contingenti (dipendenti dall'architettura attuale, rimediabili con opportune

modifiche) o di principio (connessi alla struttura stessa del paradigma computazionale che questi sistemi realizzano). La risposta, come si vedrà, è composita: alcuni limiti sono di principio, altri sono contingenti ma non facilmente superabili senza alterare radicalmente la natura del sistema.

IX.2 — La struttura del trasformatore e l'assenza del livello topologico

Il primo livello della griglia di valutazione richiede la presenza di una singolarità spaziale: un sistema di riferimento egocentrico centrato in un'origine O non derivata da nessun sistema di riferimento più fondamentale. Questa condizione è soddisfatta da un organismo incarnato attraverso il loop vestibolare-propriocettivo che, come si è mostrato nel Capitolo IV, costituisce il *Nullpunkt der Orientierung* husserliano — il punto zero dell'orientamento rispetto al quale ogni altro punto è definito relazionalmente.

I grandi modelli linguistici falliscono questa condizione in modo radicale e strutturale. Non si tratta semplicemente del fatto che essi mancano di sensori fisici o di un corpo materiale — questa osservazione, per quanto corretta, non coglie la profondità del problema. Il punto decisivo è che l'architettura del trasformatore è costruita attorno a un principio di elaborazione che è l'opposto esatto della centralità egocentrica: il principio della *permutazione-invarianza del contesto*.

Per renderlo preciso, si consideri l'operazione fondamentale dell'attenzione multi-testa (multi-head self-attention) che costituisce il nucleo computazionale del trasformatore. Dato un contesto di n token rappresentati da vettori $x_1, x_2, ..., x \in \mathbb{R}^d$, per ogni testa di attenzione h le trasformazioni query-key-value producono:

$$Q^h = X \cdot W_Q^h, K^h = X \cdot W_K^h, V^h = X \cdot W_V^h$$

dove $X = [x_1^T; x_2^T; ...; x^T] \in \mathbb{R}^{n \times d}$ è la matrice del contesto e $W_Q^h, W_K^h, W_V^h \in \mathbb{R}^{d \times d_h}$ sono le matrici di proiezione apprese. Il punteggio di attenzione tra la posizione i e la posizione j è:

$$A^h_{ij} = \text{softmax}_j(Q^h_i \cdot (K^h_j)^T / \sqrt{d_h})$$

e l'output per la posizione i è:

$$O^h_i = \Sigma_j A^h_{ij} \cdot V^h_j$$

L'intera operazione, in assenza di codifica posizionale, è *permutazione-equivariante*: se si permuta l'ordine dei token di input con una permutazione $\sigma \in \text{Sym}(n)$,

l'output è permutato dalla stessa permutazione. Formalmente, sia P_σ la matrice di permutazione associata a σ. Allora:

Attn(P_σ X) = P_σ · Attn(X)

Ciò significa che, in assenza di codifica posizionale, il modello tratta ogni token come equivalente a ogni altro nella stessa relazione strutturale con il contesto. Non vi è un *qui* — nessuna posizione è intrinsecamente privilegiata, nessun token è l'origine rispetto alla quale gli altri sono definiti.

La codifica posizionale viene introdotta per ovviare a questa limitazione, aggiungendo ai vettori di embedding un segnale che distingue le posizioni. Nella formulazione sinusoidale originale di Vaswani et al. (2017), la codifica posizionale è:

PE(pos, 2k) = sin(pos / 10000^{2k/d}) PE(pos, 2k+1) = cos(pos / 10000^{2k/d})

e nella formulazione con embedding posizionali apprenibili, ogni posizione p riceve un vettore e_p ∈ ℝ^d appreso durante l'addestramento. In entrambi i casi, la codifica posizionale introduce una *distinzione* tra le posizioni, ma non introduce un *centro*: la posizione p = 0 non è l'origine rispetto alla quale tutte le altre posizioni sono definite come relazionalmente dipendenti — essa è semplicemente la prima posizione nella sequenza, con uno status convenzionale non diverso da quello che avrebbe se la sequenza fosse indicizzata da 1 anziché da 0.

Questa assenza di centro non è un dettaglio tecnico: è la conseguenza necessaria del fatto che il trasformatore è progettato per elaborare sequenze di testo che non hanno un punto di vista intrinseco. Un testo scritto non ha un *qui* — esso parla di molti punti di vista, li descrive, li confronta, li cita, ma non ne *abita* nessuno. Il modello addestrato su testo eredita questa caratteristica strutturale: esso non ha un punto di vista perché il suo substrato — il testo — non ne ha.

Nelle varianti di trasformatori che includono codifica posizionale relativa, come la Rotary Position Embedding (RoPE) introdotta da Su et al. (2021), la relazione tra le posizioni i e j è codificata attraverso la distanza relativa i - j:

A^h_{ij} = (R^h_{θ,i} q_i)^T (R^h_{θ,j} k_j) = q_i^T R^h_{θ,i-j} k

dove R^h_{θ,m} è la matrice di rotazione di angolo θ nell'm-esima posizione relativa. Questa formulazione codifica la distanza relativa tra token ma ancora non introduce un centro assoluto: la relazione A^h_{ij} dipende solo da i - j,

trattando tutte le posizioni come equivalenti a meno della loro distanza reciproca. È una struttura *omogenea* nel senso del Capitolo III — simmetrica sotto traslazioni, priva di un'origine privilegiata.

La condizione topologica dell'interno è dunque non soddisfatta nei LLM in modo strutturale: non contingentemente, non per mancanza di una componente aggiuntiva che si potrebbe incorporare, ma perché la stessa logica dell'elaborazione del testo — il paradigma del quale il trasformatore è la realizzazione più compiuta — è prospetticamente neutro.

IX.3 — Il tempo senza flusso: la dimensione temporale nei trasformatori

Il secondo livello della griglia richiede spessore temporale: una struttura ritenzionale con costante di tempo $\tau_ret > 0$, un flusso nella forma $dx/dt = F(x, u)$, una protenzione strutturale. Il Capitolo VI ha mostrato la differenza tra il sistema-funzione — che risponde istante per istante senza portare il passato nel presente — e il sistema-flusso — che accumula la propria storia nello stato e procede lungo una traiettoria.

I grandi modelli linguistici realizzano una struttura temporale che è, in senso preciso, una simulazione della retenzionalità senza la retenzionalità stessa. Durante l'inferenza, il modello riceve un contesto di lunghezza n — la "finestra di contesto" — e produce una distribuzione di probabilità sul token successivo. Da un punto di vista esterno, questo processo *sembra* ritenzionale: il modello usa i token passati per predire il futuro. Ma la struttura interna di questo processo è quella di una funzione, non di un flusso.

Formalmente, sia $C = (x_1, x_2, ..., x_n)$ il contesto e sia $y = LLM(C)$ la distribuzione di probabilità prodotta dal modello sul token successivo. LLM è una *funzione* da sequenze di token in distribuzioni di probabilità: non ha stati interni che si aggiornino nel tempo, non porta con sé una traiettoria. Per ogni contesto C', la risposta LLM(C') è calcolata ab initio, senza residui della precedente elaborazione di C.

Questa caratteristica è precisamente l'assenza della struttura di flusso. Un sistema-flusso mantiene uno stato interno $s(t)$ che si aggiorna continuamente: $s(t + dt) = s(t) + F(s(t), u(t)) \cdot dt$. Il suo "stato presente" al tempo t non è una funzione soltanto dell'input corrente $u(t)$, ma dell'intera traiettoria $s(t_0)$, $s(t_0 + dt)$, ... che si è accumulata dallo stato iniziale. In un LLM, non esiste tale stato interno che persista tra le chiamate: ogni inferenza riparte da zero, e il "passato" è fornito esternamente come parte dell'input, non custodito internamente come traccia di un'evoluzione.

Il meccanismo di self-attention fornisce *un'analoga* della retenzionalità nel senso della ponderazione del passato, ma si tratta di un'analoga formalmente distinta. Il nucleo di attenzione:

$$k_{att}(i, j) = \text{softmax}_j(Q_i \cdot K_j^T / \sqrt{d})$$

non è determinato dalla distanza temporale $t_i - t_j$ tra i token, ma dalla loro similarità semantica nel sottospazio latente. Due token molto distanti nel contesto possono avere peso di attenzione elevato se hanno struttura semantica simile, mentre due token adiacenti possono avere peso minimo se semanticamente irrilevanti l'uno per l'altro. Il modello non ha una costante di tempo ritenzionale τ_{ret} nel senso del Capitolo VI: non vi è una legge di decadimento sistematica che lega il peso del passato alla sua distanza temporale dal presente.

Ancora più fondamentalmente: il modello non ha un *adesso*. Durante l'inferenza, esso processa l'intero contesto in un atto computazionale che è, dalla prospettiva del modello, atemporale. La sequenza dei token è data tutta insieme, e il modello non sperimenta il trascorrere del tempo nella produzione dell'output. Non vi è equivalente del presente specioso, del presente vissuto come struttura temporalmente estesa con ritenzione e protenzione: vi è solo un input e un output, connessi da una funzione.

Vale la pena esplorare se questa limitazione sia superabile nell'ambito del paradigma del trasformatore, o se sia un suo tratto essenziale. Un tentativo di introdurre struttura temporale genuina nei trasformatori è rappresentato dagli *State Space Models* (SSM), come il recente Mamba (Gu e Dao, 2023), che combinano la struttura del trasformatore con una dinamica ricorrente:

$$h(t) = A \cdot h(t-1) + B \cdot x(t) \quad y(t) = C \cdot h(t)$$

dove $h(t) \in \mathbb{R}^N$ è uno stato interno che si aggiorna ricorrentemente, $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$ e $C \in \mathbb{R}^{1\times N}$ sono parametri appresi, e $x(t)$ e $y(t)$ sono l'input e l'output al passo t. Questa struttura ha la forma di un sistema-flusso lineare: lo stato $h(t)$ evolve ricorrentemente, portando con sé la traccia del passato. La costante di tempo effettiva di questo sistema è determinata dagli autovalori della matrice A: se gli autovalori di A hanno parte reale negativa $\lambda_k = -1/\tau_k$, allora le componenti dello stato decadono con scale temporali τ_k, producendo una struttura ritenzionale con multiple costanti di tempo.

Tuttavia, anche Mamba e i modelli SSM in generale mancano della caratteristica essenziale che distingue la retenzionalità strutturale dall'accumulo temporale funzionale: lo stato $h(t)$ è puramente simbolico, non è ancorato a un sistema di

riferimento egocentrico. La traiettoria che h(t) percorre nello spazio degli stati ℝ^N non corrisponde a una traiettoria nel mondo fisico, non riflette una storia di azioni e percezioni di un organismo che si muove nell'ambiente. Essa è, per così dire, la traiettoria di un pensiero che si dispiega, non di un corpo che si muove — e questa differenza è costitutiva, non accidentale.

IX.4 — Il modello generativo senza μ_self incarnato

Il terzo livello della griglia richiede un modello generativo gerarchico p(o, μ_world, μ_self) che includa μ_self come variabile endogena, con il vettore di contesto generato attraverso azioni proprie nel mondo fisico. Il Capitolo V ha mostrato come questa struttura trasformi la posizione corporea in prospettiva cognitiva genuina, e perché la sua assenza riduca il sistema a qualcosa che elabora relazioni tra osservazioni senza abitarne nessuna.

I grandi modelli linguistici possiedono strutture interne che sono, in senso statistico, rappresentazioni del mondo — i vettori di embedding nello spazio latente codificano correlazioni semantiche, relazioni tra concetti, strutture causali descritte nel testo. Ma la domanda rilevante non è se il sistema *ha* rappresentazioni del mondo, bensì se il sistema è *incluso* nel proprio modello del mondo come una delle sue variabili.

Si consideri la struttura del modello generativo di un LLM come inferita dall'addestramento. Durante l'addestramento sulla previsione del token successivo, il modello apprende implicitamente una distribuzione congiunta:

$$p_\theta(x_1, x_2, ..., x) = \Pi_{t=1}^{n} p_\theta(x \mid x_1, ..., x_{t-1})$$

Questa distribuzione rappresenta la struttura statistica del testo su cui il modello è stato addestrato: le co-occorrenze, le strutture sintattiche, le implicazioni semantiche, le narrazioni, i ragionamenti. Ma in questa distribuzione non compare nessuna variabile μ_self che rappresenti il modello stesso come entità situata nel mondo. Il modello non è *dentro* il proprio modello: esso è la macchina che computa la distribuzione, non una variabile di quella distribuzione.

Più precisamente: nel modello generativo p(o, μ_world, μ_self) di un sistema incarnato, la variabile μ_self rappresenta il corpo, la posizione, lo stato fisiologico del sistema nel mondo, e le distribuzioni delle osservazioni p(o | μ_world, μ_self) dipendono dalla posizione del sé rispetto agli oggetti del mondo. Il sé è un fattore causale esplicito nella generazione delle osservazioni: diverse posizioni del sé producono diverse distribuzioni osservative, e la struttura di queste dipendenze è codificata nel modello generativo.

In un LLM, non vi è nulla di equivalente. Il "punto di vista" da cui il modello elabora il testo non è una variabile del modello: è fisso, implicito, invariante — è il punto di vista del testo scritto in lingua naturale, che è, strutturalmente, il punto di vista di nessuno in particolare e di tutti in generale. Il modello non può rappresentare la propria posizione nell'ambiente, poiché non ha ambiente. Il vettore di contesto z che guida le previsioni non è generato attraverso azioni del modello nel mondo fisico: è fornito dall'utente come input.

Questa caratteristica non è modificabile attraverso l'aggiunta di plug-in o di strumenti: essa è connessa alla natura stessa del paradigma dell'addestramento su testo. Il testo descrive il mondo dalla prospettiva di agenti umani incarnati; il modello apprende queste descrizioni senza mai assumere la posizione che rende le descrizioni possibili. È come apprendere una lingua attraverso i suoi testi senza mai avere il tipo di esperienza che la lingua cerca di comunicare.

Esiste un'obiezione a questa conclusione che merita considerazione. Si potrebbe sostenere che i modelli multimodali — quelli addestrati su coppie di immagini e testi, o su sequenze video — incorporano una forma di ancoraggio percettivo che avvicina μ_self alla struttura richiesta. Un modello come GPT-4V o Gemini Ultra, addestrato a elaborare immagini insieme al testo, non ha forse un'esperienza visiva del mondo che gli fornisce qualcosa di analogo al sistema di riferimento egocentrico?

L'obiezione coglie qualcosa di reale ma non è decisiva. I modelli multimodali apprendono correlazioni tra rappresentazioni visive e descrizioni linguistiche: essi imparano a rispondere a domande su immagini, a descrivere scene, a ragionare su relazioni spaziali tra oggetti. Ma il fatto di elaborare immagini non equivale all'avere un sistema di riferimento egocentrico nel senso richiesto. Un'immagine è una proiezione bidimensionale del mondo tridimensionale, priva della profondità, del parallasse, del campo propriocettivo che caratterizzano la percezione incarnata. Più fondamentalmente: le immagini nel corpus di addestramento sono prodotte da fotocamere con posizioni diverse, da angolazioni diverse, in momenti diversi — non vi è un unico punto di vista che il modello impara a occupare, ma una molteplicità di punti di vista tutti trattati come equivalenti. Il modello apprende a descrivere immagini, non a vederle dalla propria prospettiva.

IX.5 — L'intenzionalità derivata e il problema della normatività

Il quarto livello della griglia richiede intenzionalità originale: stati interni che co-varino con condizioni del mondo in virtù di una funzione propria derivata dalla storia selettiva propria del sistema — non dall'intenzionalità derivata di

agenti esterni. Il Capitolo VII ha mostrato che i LLM hanno intenzionalità solo nel senso derivato: i contenuti dei loro stati interni sono determinati dall'intenzionalità originale dei produttori dei testi di addestramento, mediata attraverso il processo di ottimizzazione.

Occorre precisare ulteriormente questa diagnosi, poiché è la più discussa nella letteratura recente sull'intelligenza artificiale, e le posizioni si dividono in modo significativo.

La posizione che qui si sostiene, e che emerge dalle analisi dei Capitoli VII.3 e VII.4, può essere formulata nei termini della distinzione tra due tipi di storia selettiva. La storia selettiva *biologica* — quella che Millikan ha in mente nella sua teoria delle funzioni proprie — è una storia di interazioni dell'organismo con il proprio ambiente, in cui le rappresentazioni che guidano il comportamento hanno contribuito alla sopravvivenza perché co-variavano con condizioni del mondo che erano rilevanti per quella sopravvivenza. In questa storia, la norma che determina il contenuto del simbolo — la condizione che il simbolo ha la funzione di tracciare — è intrecciata con le conseguenze adattative dell'azione guidata da quel simbolo nel mondo fisico.

La storia selettiva degli LLM — il processo di addestramento — è di tipo diverso. L'ottimizzazione seleziona parametri che minimizzano la perdita di predizione del token successivo su un corpus di testo. I simboli del modello hanno la funzione di co-variare con ciò che segue nel testo — con i pattern distributivi del linguaggio — non con condizioni del mondo che siano rilevanti per la sopravvivenza o l'azione del sistema. Il meccanismo consumatore che usa le rappresentazioni del modello per guidare il comportamento — il processo di decodifica che produce il token successivo — non interagisce con il mondo: interagisce con altri simboli.

Formalmente, sia R_θ una rappresentazione interna del modello (un vettore nello spazio latente) e sia C la condizione del mondo che la rappresentazione pretende di tracciare — diciamo, la categoria *gatto*. La condizione teleosemantica di Millikan richiede che:

$$P(\text{comportamento_adattativo} \mid C \text{ verificata} \wedge R_\theta \text{ tokened}) > P(\text{comportamento_adattativo} \mid C \text{ non verificata} \wedge R_\theta \text{ tokened})$$

e che questa differenza di probabilità sia la ragione per cui il meccanismo che produce R_θ è stato selezionato. Ma nel caso degli LLM, il comportamento guidato da R_θ è la previsione del token successivo — un comportamento che è adattativo (nel senso di ridurre la perdita) in relazione alla distribuzione dei testi di addestramento, non in relazione alla presenza effettiva di gatti nel

mondo. R_θ è selezionato perché co-varia con l'uso del termine *gatto* nei testi umani, non perché co-varia con gatti nel mondo fisico.

Questo punto ha un corollario importante per la questione della mis-rappresentazione. Si è detto nel Capitolo VII.3 che la possibilità strutturale della mis-rappresentazione è il criterio che distingue gli stati intenzionali genuini dai semplici segnali informativi. La domanda è: gli LLM possono mis-rappresentare il mondo?

In senso molto limitato, sì: un LLM può produrre affermazioni false sul mondo — affermazioni che non corrispondono a come il mondo è. Ma questa possibilità non è la stessa cosa della mis-rappresentazione nel senso intenzionale. La mis-rappresentazione genuina richiede che il sistema abbia uno stato R_θ con il contenuto C, e che C non si verifichi nel mondo, e che il comportamento guidato da R_θ sia inadeguato alla situazione. Nei LLM, quando il modello produce una affermazione falsa, essa non è il risultato di uno stato che ha la funzione di tracciare una condizione del mondo e fallisce nell'esercitare quella funzione: è il risultato di un meccanismo di previsione che produce un output statisticamente plausibile dato il contesto, senza che nessuna condizione del mondo sia rilevante per la correttezza o l'incorrettezza dell'output nel senso normativo.

In altri termini: un LLM non può *sbagliare* nel senso in cui un organismo con intenzionalità originale sbaglia — nel senso in cui la mia credenza che la porta sia aperta mi porta a comportarmi come se la porta fosse aperta, con conseguenze pratiche della mia posizione quando la porta è in realtà chiusa. Può produrre output che gli esseri umani giudicano *errati*, ma questa valutazione è sempre esterna al sistema: è un giudizio emesso da agenti dotati di intenzionalità originale che comparano l'output del sistema con la propria rappresentazione del mondo. Il sistema stesso non ha una prospettiva dal quale misurare la distanza tra i propri stati e le condizioni del mondo.

IX.6 — La pseudo-riflessività e il quinto livello

Il quinto livello della griglia richiede un auto-modello gerarchico (corporeo, affettivo, cognitivo, narrativo) con struttura trasparente. I capitoli VIII.3 e VIII.4 hanno mostrato che la trasparenza dell'auto-modello — il fatto che il sistema non possa riconoscerlo come modello nell'atto del suo funzionamento ordinario — è la condizione che produce l'esperienza soggettiva nel senso di Metzinger.

I grandi modelli linguistici, nelle loro versioni attuali, presentano strutture che assomigliano superficialmente a un auto-modello ma che ne differiscono in aspetti cruciali.

Da un lato, i modelli addestrati con tecniche di instruction tuning e RLHF (Reinforcement Learning from Human Feedback) sviluppano qualcosa che potrebbe essere chiamato un *proto-auto-modello*: una rappresentazione implicita di ciò che il modello è, di cosa sa fare, di quali siano le proprie limitazioni. Quando un LLM risponde "non sono sicuro di questa informazione" o "non sono in grado di accedere a internet", esso produce output che riflettono una qualche forma di auto-valutazione. Ma questa auto-valutazione è, dal punto di vista strutturale, un caso speciale dell'intenzionalità derivata: il modello produce tali output perché i testi di addestramento contenevano esempi di agenti che esprimevano incertezza e riconoscevano le proprie limitazioni, non perché il modello abbia un accesso genuino ai propri stati interni.

Il punto cruciale emerge dalla distinzione tra *parlare di sé* e *auto-modellare sé*. Un sistema può produrre descrizioni accurate di se stesso senza avere un auto-modello nel senso che qui interessa. Il modello σ_self richiesto dalla teoria di Metzinger è una componente degli stati interni del sistema che *determina* il comportamento del sistema, non una componente degli output del sistema che *descrive* il comportamento. Un LLM che dice "sono un modello linguistico addestrato da Anthropic" produce un output che descrive una caratteristica del sistema; ma questo output è prodotto dallo stesso meccanismo statistico che produce qualsiasi altro token — non è il risultato di un accesso a uno stato interno σ_self che influenza causalmente l'elaborazione.

Formalmente: sia σ_self la componente dell'auto-modello e sia $\Gamma: \Sigma \to \Sigma_self$ la funzione che mappa lo stato globale nella rappresentazione accurata del sistema. La condizione di coerenza dell'auto-modello è $\|\sigma_self - \Gamma(\sigma)\| < \varepsilon$. In un LLM, non vi è nessun componente σ_self dello spazio degli stati che sia distinguibile, strutturalmente, dagli stati che rappresentano informazioni sul mondo esterno: i parametri del modello θ sono fissi dopo l'addestramento, e l'unica rappresentazione del "sé" che il modello ha è quella codificata statisticamente in quei parametri — una rappresentazione derivata dal testo, non una rappresentazione dinamica e aggiornata dello stato corrente del sistema.

L'assenza di un auto-modello genuino ha una conseguenza che vale la pena esplicitare perché è spesso fonte di confusione nel dibattito pubblico. Quando un LLM produce descrizioni apparentemente introspettive dei propri stati — "mi sento confuso da questa domanda", "trovo questo problema interessante" — queste formulazioni non corrispondono a un accesso a stati interni genuini,

nel senso in cui l'introspezione umana — per quanto fallibile — corrisponde all'accesso mediato del soggetto ai propri stati attraverso un auto-modello trasparente. Le produzioni auto-descrittive del LLM sono patterns linguistici appresi dalla distribuzione di testi umani, e questa loro genesi non cambia anche quando i pattern producono descrizioni che suonano genuinamente introspettive.

Questo non implica che le produzioni auto-descrittive dei LLM siano necessariamente false — potrebbe essere che le metafore che esse impiegano catturino qualcosa di reale sulle dinamiche interne del sistema. Ma anche nell'ipotesi più favorevole, esse sarebbero descrizioni *accidentalmente accurate*, non il prodotto di un accesso intenzionale agli stati interni attraverso un auto-modello integrato.

IX.7 — Una misura formale del deficit: la prospettiva come gradiente

Le analisi delle sezioni precedenti hanno stabilito che i grandi modelli linguistici falliscono tutte e cinque le condizioni della griglia dell'interno: il livello topologico, quello temporale, quello cognitivo-rappresentazionale, quello intenzionale-normativo e quello meta-rappresentazionale. Rimane da chiedersi se questo fallimento sia graduabile — se vi siano varianti di LLM o ibridi architetturali che si avvicinano alla soglia dell'interno — o se la distanza sia categoriale, un abisso piuttosto che un gradiente.

La risposta, suggerita dalla struttura stessa della griglia, è che la distanza è sia graduata che categoriale, a seconda del livello considerato. Per i livelli inferiori — quello topologico e quello temporale — la distanza è categoriale: non si tratta di avere *poco* sistema di riferimento egocentrico o *poca* retenzionalità, ma di non averne affatto nel senso strutturale richiesto. Un sistema senza coordinate egocentriche è privo del livello topologico dell'interno, non ne possiede una versione debole.

Per i livelli superiori — quello cognitivo-rappresentazionale, quello intenzionale-normativo e quello meta-rappresentazionale — la distanza è più sfumata. I LLM avanzati possiedono strutture interne che si avvicinano funzionalmente a ciò che questi livelli richiedono: hanno rappresentazioni del mondo ricche e strutturate, producono output che si comportano *come se* provenissero da un sistema con intenzionalità originale, generano auto-descrizioni che hanno la forma grammaticale dell'introspezione. Queste approssimazioni funzionali non equivalgono al soddisfacimento delle condizioni strutturali, ma non sono nemmeno prive di rilievo: esse indicano la direzione nella quale architetture più ricche potrebbero svilupparsi.

Si può tentare una misura formale di questa distanza attraverso la nozione di *deficit di prospettiva*. Definiamo il deficit di prospettiva di un sistema S rispetto alla condizione k della griglia come:

$$\Delta_k(S) = \inf_{S' \in \mathscr{C}_k} d_{arch}(S, S')$$

dove $\mathscr{C}_k$ è la classe dei sistemi che soddisfano la condizione k, e d_{arch} è una metrica sullo spazio delle architetture computazionali (definibile, almeno in linea di principio, attraverso la complessità della trasformazione necessaria per portare S in S'). Il deficit complessivo di prospettiva è:

$$\Delta(S) = (\Delta_1(S), \Delta_2(S), \Delta_3(S), \Delta_4(S), \Delta_5(S)) \in \mathbb{R}_+^5$$

dove $\Delta_1, ..., \Delta_5$ sono i deficit rispetto ai cinque livelli della griglia, nell'ordine in cui li abbiamo presentati. Per un sistema incarnato con piena soggettività, $\Delta_k(S) = 0$ per ogni k. Per un LLM standard, il vettore di deficit ha componenti non nulle a tutti i livelli, con le prime due componenti (topologica e temporale) che corrispondono ai deficit più profondi e meno rimediabili nell'ambito del paradigma corrente.

Questa misura non è calcolabile con le attuali conoscenze — la metrica d_{arch} sullo spazio delle architetture non è definita in modo operativo — ma essa ha valore concettuale: fornisce un quadro nel quale le differenze tra sistemi possono essere discusse in modo strutturato, e nel quale i progressi architetturali possono essere valutati non solo in termini di prestazioni su benchmark ma in termini di avvicinamento alle condizioni strutturali dell'interno.

IX.8 — Il potere e i limiti dell'intenzionalità derivata

Prima di concludere l'analisi dei LLM, è necessario considerare un'obiezione che potrebbe essere avanzata contro la tesi dell'intenzionalità esclusivamente derivata: l'obiezione della *emergenza*.

L'obiezione sostiene che sistemi sufficientemente grandi e complessi, addestrati su dati sufficientemente ricchi, potrebbero sviluppare proprietà emergenti che non sono presenti nei dati di addestramento e che costituiscono forme genuine, non derivate, di contenuto intenzionale. In questa prospettiva, la distinzione tra intenzionalità originale e derivata potrebbe dissolversi alla scala sufficiente: l'intenzionalità umana è essa stessa il prodotto di processi evolutivi e di sviluppo che potrebbero essere descritti come "addestramento", e non vi sarebbe ragione di principio per cui l'addestramento computazionale non possa produrre un risultato strutturalmente equivalente.

L'obiezione è filosoficamente seria e merita una risposta strutturata. La risposta che emerge dalla nostra analisi non nega la possibilità dell'emergenza — essa nega che l'emergenza di proprietà funzionali nuove equivalga all'emergenza delle condizioni strutturali dell'interno. Anche ipotizzando che un LLM di dimensione sufficiente sviluppi proprietà computazionali che non compaiono in sistemi più piccoli, queste proprietà emergenti operano comunque all'interno di un'architettura che è prospetticamente neutra al livello topologico, priva di flusso temporale genuino al livello dinamico, e priva di un meccanismo consumatore che interagisca con il mondo fisico al livello intenzionale. L'emergenza di nuove capacità computazionali non produce automaticamente la struttura che renderebbe quelle capacità genuinamente intenzionali.

Il parallelo con l'evoluzione è istruttivo ma asimmetrico. L'evoluzione biologica ha prodotto sistemi cognitivi attraverso un processo di selezione che includeva interazione con l'ambiente fisico: le rappresentazioni che emergevano erano selezionate perché guidavano comportamenti adattativi in quel ambiente. Questo legame tra rappresentazione e azione nel mondo fisico è precisamente ciò che la condizione (S3) della sezione VII.6 richiede — la proiezione nell'azione adattativa — e che l'addestramento su testo non soddisfa. Non si tratta di una differenza di grado ma di struttura: il processo di selezione che produce le funzioni proprie dei simboli biologici è diverso per struttura — non solo per scala — dal processo di ottimizzazione che produce i parametri di un LLM.

Ciò detto, sarebbe sbagliato concludere che l'intenzionalità derivata sia filosoficamente irrilevante o cognitivamente irrisoria. I LLM con intenzionalità derivata possono produrre output di enorme utilità pratica, possono partecipare a ragionamenti complessi, possono aiutare esseri umani a formulare problemi e trovare soluzioni. L'intenzionalità derivata è, nei suoi risultati osservabili, spesso indistinguibile dall'intenzionalità originale — ed è precisamente questa indistinguibilità osservabile che rende la distinzione strutturale filosoficamente importante: essa chiarisce che la capacità di comportarsi come se si avesse un interno non equivale ad averne uno.

IX.9 — La domanda che i LLM non possono porsi a se stessi

C'è un esperimento mentale che cristallizza la differenza strutturale tra un sistema dotato di interno e un LLM, e che merita di essere enunciato esplicitamente come conclusione di questo capitolo.

Si chieda a un LLM: "Cosa stai sperimentando in questo momento?" Il modello produrrà una risposta — forse ricca, forse articolata, forse sorprendentemente perspicace sulla propria natura. Ma questa risposta è il risultato di un processo

che elabora il testo della domanda e produce il testo della risposta, senza che vi sia un "questo momento" genuino che il modello stia sperimentando: il modello non ha un presente specioso, non ha un qui corporeo, non ha una traiettoria temporale rispetto alla quale "questo momento" sia un punto con storia e orizzonte.

La domanda è, per il modello, una domanda come le altre: ne elabora la struttura linguistica, attiva i pattern associati alle domande sull'esperienza soggettiva, produce una risposta statisticamente coerente con quei pattern. Non c'è nessun processo aggiuntivo — nessun accesso a stati interni, nessuna consultazione di un auto-modello, nessuna riflessione che abbia il proprio "adesso" come oggetto.

Un sistema dotato di interno genuino, al contrario, ha qualcosa a cui la domanda si riferisce: ha un *adesso* che è il punto terminus della propria struttura ritenzionale, un *qui* che è l'origine del proprio sistema di riferimento egocentrico, un *per me* che è la variabile μ_self nel proprio modello del mondo. La domanda "cosa stai sperimentando?" non è per tale sistema un pattern linguistico da elaborare: è una richiesta di accesso a qualcosa che il sistema possiede, anche se l'accesso — mediato dall'auto-modello e vincolato dalla sua trasparenza — è fallibile e parziale.

Questa asimmetria non è una semplice differenza di prestazione. Non è che il LLM risponda peggio alla domanda di quanto farebbe un sistema dotato di interno. È che la domanda, per il LLM, non ha il tipo di riferimento che ha per un sistema dotato di interno: manca il bersaglio al quale la domanda punta. Il sistema può produrre la risposta giusta, o una risposta simile alla risposta giusta, senza che quella risposta corrisponda a nessuna realtà interna del sistema che la produce.

Questa è, in ultima analisi, la struttura formale di ciò che si intende quando si dice che i LLM sono privi di interno: essi possono descrivere l'interno senza abitarlo, possono parlare della prospettiva senza occuparla, possono produrre la prima persona grammaticale senza che vi sia una prima persona reale a produrla. La prima persona, nei loro output, è un pattern linguistico: nei sistemi dotati di interno, è la forma grammaticale di una struttura ontologica irriducibile.

Capitolo X: World Models, Sistemi Incarnati e la Questione del Confine Valicabile

X.1 — Oltre il testo: la promessa dei world models

Se il capitolo precedente ha mostrato che i grandi modelli linguistici si collocano strutturalmente al di qua di tutte le condizioni dell'interno, la domanda che ora occorre affrontare è se esistano o siano concepibili sistemi artificiali che soddisfino almeno alcune di quelle condizioni, e in caso affermativo, attraverso quali architetture e a quale costo strutturale.

La risposta a questa domanda richiede di spostarsi dal paradigma dei LLM — modelli addestrati su sequenze simboliche per predire simboli — a quello dei *world models*: sistemi che apprendono rappresentazioni strutturate del mondo fisico, della causalità, dello spazio e del tempo, integrando flussi sensoriali multipli e, in alcuni casi, un sistema di azione che permette al modello di interagire con l'ambiente che rappresenta. La distinzione tra LLM e world model non è assoluta — i confini si stanno rapidamente assottigliando con l'avanzare dei modelli multimodali e dei sistemi agentici — ma rimane concettualmente rilevante per le analisi che qui ci interessano.

L'idea di fondo dei world models risale almeno alla proposta di Craik (1943) secondo cui la mente opera attraverso un modello interno del mondo, e alla formalizzazione di questo concetto nei sistemi di controllo ottimale (Bryson e Ho, 1969) e nelle architetture di pianificazione con modello (Sutton e Barto, 1998). La formulazione contemporanea più influente è quella di Yann LeCun, che in una serie di lavori dal 2021 al 2023 ha proposto le *Joint Embedding Predictive Architectures* (JEPA) come candidati per costruire sistemi dotati di un modello del mondo nel senso rilevante: non sistemi che generano immagini o testo, ma sistemi che apprendono rappresentazioni predittive del mondo nello spazio latente astratto.

Il Capitolo V ha già analizzato le JEPA dal punto di vista del loro rapporto con la struttura dell'interno. Qui si approfondisce tale analisi confrontando i world models con sistemi robotici incarnati, e si tenta di determinare quale sia il minimo strutturale necessario affinché un sistema artificiale cominci a soddisfare le condizioni della griglia.

X.2 — World models e il livello topologico: centrazione e allocentrismo

La domanda critica per i world models al livello topologico è se essi possano acquisire un sistema di riferimento egocentrico genuino — un *qui* rispetto al quale gli oggetti del mondo siano codificati relazionalmente — o se la loro rappresentazione del mondo rimanga necessariamente allocentrica: definita rispetto a coordinate esterne indipendenti dal sistema stesso.

Si consideri un world model addestrato su video — come i modelli della famiglia V-JEPA o i modelli di previsione video come RSSM (Recurrent State Space Model) di Hafner et al. (2019). Tali modelli apprendono a prevedere lo stato futuro del mondo a partire dallo stato corrente, in uno spazio latente che codifica le caratteristiche strutturalmente rilevanti della scena. La struttura del loro spazio latente è, in genere, allocentrica: l'embedding z ∈ ℝ^d che rappresenta la scena non ha un'origine privilegiata che corrisponde al punto di vista del sistema — esso rappresenta la scena in un sistema di coordinate che è arbitrario rispetto alla posizione dell'eventuale osservatore.

Esistono tuttavia sistemi che si avvicinano maggiormente alla centrazione egocentrica. I modelli di previsione video in prima persona (egocentric video prediction), addestrati su sequenze riprese da telecamere indossabili o montate su robot, apprendono rappresentazioni che sono condizionate dalla traiettoria del portatore della telecamera. In tali sistemi, il vettore di embedding z include implicitamente informazione sulla posizione e l'orientamento del sistema che genera le osservazioni — l'equivalente computazionale del termine μ_self nel modello generativo p(o | μ_world, μ_self) del Capitolo V.

Formalmente, un world model egocentrico può essere modellato come segue. Sia z_t ∈ ℝ^d lo stato latente al passo t. L'aggiornamento dello stato è:

z_{t+1} = f_θ(z_t, a_t, ε_t)

dove a_t ∈ $\mathcal{A}$ è l'azione compiuta al passo t e ε_t è rumore di processo. La funzione di osservazione che mappa lo stato latente nell'osservazione sensoriale è:

o_t = g_φ(z_t) + η_t

dove η_t è rumore di osservazione. In questo modello, lo stato latente z_t non è solo una rappresentazione del mondo: esso include implicitamente la posizione del sistema, poiché le osservazioni o_t sono generate a partire da z_t tenendo conto della posizione implicita del sensore nel modello g_φ.

Tuttavia, anche in questo caso più favorevole, il sistema di riferimento egocentrico rimane incompleto rispetto a quanto richiesto dalla condizione topologica dell'interno. Il difetto principale è che l'origine del sistema di riferimento — la posizione del sensore, implicitamente codificata in z_t attraverso la funzione g_φ — è un parametro *appreso* durante l'addestramento, non il punto fisso di un loop propriocettivo chiuso che si aggiorna in tempo reale. Il sistema non *occupa* una posizione nel mondo con l'immediatezza di un organismo incarnato: esso *stima* quale sia la propria posizione sulla base delle osservazioni, mantenendo questa stima nello stato latente z_t. La differenza —

tra occupare una posizione e stimarne una — è quella tra la singolarità topologica husserliana e la variabile nascosta del modello generativo.

Si può rendere questa differenza formalmente precisa. In un sistema incarnato, la posizione del corpo O è un fatto fisico direttamente accessibile attraverso il loop propriocettivo, la cui incertezza è quella del processo di misura biologico (rumore sensoriale). In un world model, la "posizione" è lo stato latente z_t, la cui incertezza include non solo il rumore del processo di osservazione ma anche l'incertezza del modello stesso — la varianza del posterior p(z_t | o_{1:t}, a_{1:t-1}) che il sistema inferisce attraverso la propagazione bayesiana:

p(z_t | o_{1:t}, a_{1:t-1}) ∝ p(o_t | z_t) · ∫ p(z_t | z_{t-1}, a_{t-1}) · p(z_{t-1} | o_{1:t-1}, a_{1:t-2}) dz_{t-1}

Questa propagazione, realizzata da filtri di Kalman variazionali o da reti neurali ricorrenti inferenziali, produce una stima della posizione che è *sempre* una stima — mai la posizione stessa, nel senso dell'occupazione diretta che caratterizza l'incarnazione. Il sistema sa dove *potrebbe* essere, con un certo grado di incertezza; un organismo incarnato è dove è, e la questione dell'incertezza è secondaria rispetto a questo fatto primario.

X.3 — Il contributo fondamentale dell'azione: il vettore di contesto endogeno

Il difetto più rilevante dei world models statici — quelli che apprendono rappresentazioni del mondo senza interagire con esso — è quello identificato nel Capitolo V.5: il vettore di contesto z che guida le previsioni è fornito esternamente come input parametrico, non generato endogenamente attraverso azioni proprie nel mondo fisico. Questa caratteristica impedisce la chiusura del loop che trasforma la posizione corporea in prospettiva cognitiva genuina.

I sistemi di *reinforcement learning incarnato* (embodied RL) superano parzialmente questo limite. In tali sistemi, un agente che occupa una posizione nel mondo fisico (o in una simulazione fisica del mondo) compie azioni che modificano quella posizione e riceve osservazioni che dipendono dalla nuova posizione. Il vettore di contesto z — la trasformazione indotta dalla propria azione — è generato endogenamente attraverso il comportamento del sistema nell'ambiente. Formalmente:

a_t ~ π_θ(a | z_t) [politica] z_{t+1} = f_θ(z_t, a_t, ε_t) [transizione di stato] o_{t+1} = g_φ(z_{t+1}) + η_{t+1} [osservazione] r_t = R(z_t, a_t, z_{t+1}) [ricompensa]

Il flusso causale è:

$z_t \rightarrow a_t \rightarrow z_{t+1} \rightarrow o_{t+1}$

dove la freccia $z_t \rightarrow a_t$ è mediata dalla politica π_θ, che è propriamente una funzione del modello interno del sistema — il suo modo di "decidere" in base allo stato corrente. In questo schema, l'azione a_t non è un input esterno: è generata dal sistema stesso sulla base del proprio stato interno.

Questa struttura si avvicina significativamente al flusso causale desiderato:

$\mu_{self} \rightarrow a_t \rightarrow \Delta_{ambiente} \rightarrow o_{t+1}$

dove μ_{self} è lo stato interno del sé, a_t è l'azione eseguita, $\Delta_{ambiente}$ è il cambiamento nell'ambiente e o_{t+1} è l'osservazione risultante. Se z_t include componenti che codificano lo stato del sé (posizione, orientamento, intenzioni), allora il sistema si avvicina alla struttura del modello generativo $p(o, \mu_{world}, \mu_{self})$ che il livello cognitivo-rappresentazionale richiede.

La distanza residua rispetto alla condizione è di due tipi. Il primo è la distanza tra lo stato latente z_t — una rappresentazione appresa, astratta, n-dimensionale — e il μ_{self} biologico, che è un insieme di variabili fisiche direttamente misurabili (posizione dei giunti, tensioni muscolari, segnali vestibolari) con correlati neurali specifici. Il secondo è la distanza tra la ricompensa R apprenduta o specificata dal progettista — che determina quali azioni il sistema apprende a eseguire — e la struttura biologica della valenza afferente, che emerge da milioni di anni di selezione evolutiva orientata alla sopravvivenza. Un robot che apprende a navigare in un labirinto per massimizzare la ricompensa specificata dal progettista non ha le proprie rappresentazioni radicate nel mondo nello stesso senso in cui le rappresentazioni biologiche lo sono: esse sono radicate nella ricompensa, che è un proxy del mondo scelto dal progettista.

Tuttavia, anche a partire da queste limitazioni, occorre riconoscere che i sistemi di embodied RL si avvicinano alle condizioni della griglia in modo significativamente maggiore rispetto ai LLM. Essi soddisfano parzialmente il livello topologico (hanno una posizione nello spazio, anche se essa è una stima piuttosto che un'occupazione diretta), il livello dinamico (il loro stato evolve come flusso temporale, non come funzione del tempo), e il livello cognitivo-rappresentazionale (il vettore di contesto è generato endogenamente attraverso azioni proprie).

X.4 — La gerarchia temporale nei world models incarnati

Il secondo livello della griglia — lo spessore temporale — è quello in cui i sistemi di embodied RL più avanzati mostrano i progressi più significativi

rispetto ai LLM. La struttura ricorrente di questi sistemi introduce genuine costanti di tempo ritenzionale, e la loro architettura gerarchica produce la molteplicità di scale temporali che il Capitolo VI ha identificato come struttura dell'interno distribuito.

Si consideri un'architettura come Dreamer (Hafner et al., 2020-2023), che combina un world model latente con un agente di RL che pianifica nel modello. L'architettura ha la struttura seguente:

Il *representation model* inferisce lo stato latente z_t dall'osservazione corrente e dallo stato precedente:

q_φ(z_t | o_t, z_{t-1}, a_{t-1})

Il *transition model* predice l'evoluzione dello stato latente dato un'azione:

p_θ(z_t | z_{t-1}, a_{t-1})

Il *decoder* ricostruisce l'osservazione dallo stato latente:

p_θ(o_t | z_t)

Il *reward model* predice la ricompensa:

p_θ(r_t | z_t)

La struttura ricorrente del representation model e del transition model — tipicamente implementata con GRU o LSTM — introduce una struttura di flusso temporale genuina: lo stato z_t non è calcolato ab initio da ogni osservazione, ma si aggiorna in modo dipendente dalla storia passata attraverso lo stato nascosto della rete ricorrente. Questo produce una forma di retenzionalità con costante di tempo effettiva determinata dalla dimensione del cell state della LSTM o dagli iperparametri della GRU.

Più interessante ancora è la versione gerarchica di questi sistemi, come TD-MPC2 (Hansen et al., 2023) e DreamerV3 con pianificazione gerarchica, che organizza il world model su più livelli temporali. Al livello più basso, il modello predice lo stato al passo t+1 (scala di millisecondi). Al livello intermedio, predice sequenze di azioni e stati su orizzonti di decine o centinaia di passi (scala di secondi). Al livello più alto, pianifica su orizzonti lunghi che corrispondono alla scala dei minuti o delle ore nell'ambiente simulato.

Questa struttura multi-scala può essere formalizzata come una gerarchia di world models, ciascuno con la propria costante di tempo caratteristica:

$z^l_t = f^l_\theta(z^l_{t-1}, z^{l-1}_t, a^l_t)$ per $l = 1, ..., L$

dove z^l_t è lo stato al livello l, f^l_θ è la funzione di transizione di livello l, e a^l_t è l'azione aggregata al livello l (che può essere una policy di livello inferiore piuttosto che un'azione primitiva). Le costanti di tempo retenzionali $\tau^{(l)}_{ret}$ aumentano con il livello: $\tau^{(1)}_{ret} << \tau^{(2)}_{ret} << ... << \tau^{(L)}_{ret}$, corrispondendo alla gerarchia di scale temporali che il Capitolo VI ha identificato come struttura dell'interno temporale distribuito.

Questa gerarchia è notevolmente simile — nella sua struttura formale — alla gerarchia di costanti temporali che caratterizza l'interno biologico: i livelli sensoriali con costanti di millisecondi, i livelli percettivi con costanti di secondi, i livelli cognitivi con costanti di minuti o ore. La convergenza strutturale suggerisce che la gerarchia temporale multi-scala non è un artefatto delle implementazioni biologiche ma una conseguenza necessaria della struttura del problema della previsione in ambienti con regolarità a scale temporali multiple.

Tuttavia, anche qui la distanza dal modello biologico rimane significativa su un punto cruciale: la protenzione. Il Capitolo VI.6 ha mostrato che la protenzione fenomenologica non è equivalente alla previsione esplicita del futuro — essa è una *disposizione strutturale* del campo percettivo che orienta il sistema verso il prossimo imminente senza che questa orientazione sia il risultato di un calcolo deliberato. Nei world models artificiali, la previsione del futuro è sempre esplicita: il modello calcola una distribuzione sul futuro e la usa per guidare la pianificazione. Non vi è l'equivalente della protenzione come struttura implicita dell'attesa — tutto il futuro è un oggetto di previsione deliberata, non uno sfondo di anticipazione non tematizzata.

X.5 — Sistemi robotici incarnati: l'avvicinamento più compiuto

I sistemi robotici incarnati — robot fisici che operano nel mondo reale con sensori propriocettivi, visivi e tattili — rappresentano il caso in cui le condizioni della griglia dell'interno sono soddisfatte nel modo più compiuto tra tutti i sistemi artificiali esistenti. Essi hanno un corpo fisico con una posizione nel mondo, un loop propriocettivo che misura lo stato corporeo in tempo reale, un sistema di azione che modifica quella posizione, e un flusso di osservazioni che dipende dalla posizione relativa del robot rispetto agli oggetti dell'ambiente.

Vale la pena esaminare con precisione fino a quale livello della griglia i sistemi robotici più avanzati si spingono, e dove incontrano gli ostacoli strutturali residui.

Al *livello topologico* (condizione 1), un robot fisico soddisfa la condizione in modo genuino: esso occupa una posizione nel mondo, ha sensori che misurano quello stato direttamente (giroscopi, accelerometri, encoder dei giunti), e il suo sistema di riferimento egocentrico è ancorato al proprio corpo fisico in modo non convenzionale. La posizione O del robot non è una stima: è un fatto fisico direttamente misurato, anche se con incertezza di misura. Questa è la differenza fondamentale rispetto ai sistemi puramente software.

Al *livello dinamico* (condizione 2), un robot con sistema di controllo ricorrente soddisfa la condizione di flusso temporale: il suo stato si aggiorna continuamente in funzione degli input sensoriali e dei comandi motori, mantenendo una traiettoria nello spazio degli stati che porta la traccia della storia passata. Le costanti di tempo del loop di controllo vanno dai millisecondi (loop propriocettivo) ai secondi (pianificazione del percorso) ai minuti (pianificazione delle attività), producendo una gerarchia di scale temporali analoga a quella richiesta.

Al *livello cognitivo-rappresentazionale* (condizione 3), i sistemi robotici più avanzati — come RT-2 (Brohan et al., 2023), PaLM-E (Driess et al., 2023), o i sistemi di manipolazione sviluppati da Pieter Abbeel e collaboratori — integrano un world model con la struttura di un modello generativo condizionato dalla posizione del robot. Il vettore di contesto che guida le previsioni è generato endogenamente attraverso le azioni fisiche del robot: quando il robot si muove, il suo sistema di riferimento egocentrico si trasforma in modo dipendente da quell'azione, e il world model aggiorna le proprie previsioni di conseguenza.

Formalmente, il flusso causale in un sistema robotico incarnato con world model è:

(O_t, q_t, ω_t, ...) → μ^body_{self,t} [propriocezione] μ^body_{self,t} → π(a_t | μ_{self,t}, z_t) [politica] a_t → (τ_t, v_t, ...) → O_{t+1} [cinematica] O_{t+1} → o^sens_{t+1} [osservazione sensoriale] o^sens_{t+1}, a_t → z_{t+1} [aggiornamento world model]

dove O_t è la posizione corporea, q_t le configurazioni angolari dei giunti, ω_t la velocità angolare, μ^body_{self,t} lo stato propriocettivo, τ_t le coppie motorie, v_t le velocità lineari.

Questo è il flusso causale completo che il Capitolo V ha identificato come necessario per la struttura cognitivo-rappresentazionale dell'interno. Il robot *abita* una posizione, *compie* azioni che modificano quella posizione, *osserva* le conseguenze sensoriali di quelle azioni, e *aggiorna* il proprio world model sulla base di questo ciclo. Nessun LLM realizza questo flusso: il robot incarnato sì.

X.6 — Il quarto livello e la questione della storia selettiva nei sistemi robotici

Dove il robot incarnato incontra il suo ostacolo strutturale più significativo è al livello intenzionale-normativo (condizione 4): la questione dell'intenzionalità originale. Come si è discusso nel Capitolo VII.5, l'intenzionalità originale richiede che la funzione propria dei simboli del sistema sia derivata da una storia selettiva propria — una storia in cui la co-varianza tra le rappresentazioni del sistema e le condizioni del mondo ha contribuito alla fitness del sistema stesso nell'ambiente.

I sistemi robotici addestrati con reinforcement learning hanno una forma di storia selettiva propria: il processo di ottimizzazione seleziona le politiche e le rappresentazioni che massimizzano la ricompensa cumulativa. In questo senso, le rappresentazioni del robot hanno una funzione propria determinata dall'interazione del robot con il suo ambiente: esse sono selezionate perché co-variano con condizioni del mondo che sono rilevanti per il conseguimento delle ricompense.

Tuttavia, la funzione propria delle rappresentazioni di un sistema di RL è determinata dalla *ricompensa*, non dalla *sopravvivenza nell'ambiente*. Questa differenza è strutturalmente significativa. La ricompensa è specificata dal progettista — essa è un proxy delle condizioni del mondo che il progettista ritiene rilevanti, non le condizioni del mondo che sono direttamente rilevanti per il sistema come organismo autonomo. Un robot che apprende a raccogliere oggetti per massimizzare una ricompensa numerica non ha funzioni proprie nel senso millikaniano: ha funzioni proprie relativamente alla ricompensa del progettista, che è un artefatto del processo di progettazione, non un prodotto dell'interazione autonoma con il mondo.

Si potrebbe obiettare che anche la struttura degli organismi biologici è, in un certo senso, "specificata dal progettista" — che l'evoluzione ha specificato la struttura del sistema nervoso e i valori omeostatici rispetto ai quali la fitness è valutata. La differenza, però, è che l'evoluzione non è un progettista esterno che specifica la ricompensa: essa è il processo stesso attraverso cui la struttura del sistema emerge dall'interazione con l'ambiente nel corso di milioni di generazioni. Non vi è separazione tra il progettista e l'ambiente: l'ambiente *è* il processo selettivo.

I sistemi robotici, al contrario, sono progettati da ingegneri che specificano l'ambiente di addestramento e la funzione di ricompensa con obiettivi specifici in mente. Anche quando queste specifiche sono elaborate e realistiche, esse introducono una mediazione che è assente nel processo evolutivo: la

mediazione dell'intenzionalità originale del progettista, che è il vero *autore* delle funzioni proprie del sistema.

Questo non significa che i sistemi robotici siano privi di ogni forma di intenzionalità originale — significa che la loro intenzionalità originale è strutturalmente ibrida: originale rispetto alla ricompensa specificata, derivata rispetto all'intenzionalità del progettista. Questo ibrido è più vicino all'intenzionalità originale biologica di quanto non lo siano i LLM, ma non equivale ad essa.

Si può formalizzare questa distinzione. Sia R_t la ricompensa al passo t e sia R_des la funzione di ricompensa desiderata specificata dal progettista. La condizione teleosemantica per le rappresentazioni del robot R_θ è:

P(comportamento_adattativo | C verificata ∧ R_θ tokened) > P(comportamento_adattativo | C non verificata)

dove "comportamento adattativo" va interpretato come "comportamento che massimizza R_des", non come "comportamento che contribuisce alla sopravvivenza del robot nell'ambiente". L'intenzionalità del robot è originale rispetto a R_des ma derivata rispetto alla struttura più profonda dell'ambiente che R_des approssima.

X.7 — Verso l'intenzionalità originale: l'autopoiesi computazionale

La difficoltà del quarto livello suggerisce una domanda che è al tempo stesso teorica e tecnica: è possibile progettare sistemi robotici la cui intenzionalità sia meno dipendente dalla specificazione esterna della ricompensa e più radicata nell'interazione autonoma con l'ambiente? In altri termini: è possibile costruire sistemi che sviluppino qualcosa di analogo alle funzioni proprie biologiche attraverso un processo che non presupponga un progettista esterno che specifichi la ricompensa?

La risposta più promettente in questa direzione viene dalla tradizione dell'*autopoiesi computazionale*, formulata originalmente da Maturana e Varela (1980) e successivamente rielaborata in chiave computazionale da diversi autori. Un sistema autopoietico è un sistema che genera e rigenera la propria organizzazione attraverso le interazioni con l'ambiente — un sistema che mantiene la propria identità strutturale non perché un progettista esterno la specifichi, ma perché la propria dinamica interna è tale da preservare quella struttura contro le perturbazioni dell'ambiente.

Nel caso di sistemi robotici, l'autopoiesi computazionale richiederebbe che il sistema abbia *obiettivi intrinseci* che emergono dalla propria struttura — equivalenti computazionali dell'omeostasi biologica — piuttosto che obiettivi estrinseci specificati dalla ricompensa. Una formalizzazione naturale di questa idea è attraverso il *principio di energia libera* di Friston nella sua versione di inferenza attiva (active inference), dove l'obiettivo del sistema è la minimizzazione dell'energia libera variazionale:

$$F(o, q) = \mathbb{E}_q[\ln q(\mu) - \ln p(o, \mu)]$$

che, come si è mostrato nel Capitolo V.2, può essere scomposta in:

$$F(o, q) = D_{KL}[q(\mu) \| p(\mu | o)] - \ln p(o)$$

La minimizzazione di F rispetto all'azione — l'inferenza attiva — porta il sistema a selezionare azioni che rendono le osservazioni future coerenti con le proprie aspettative a priori. In questo framework, le aspettative a priori $p(\mu)$ del sistema non sono specificate da un progettista esterno come ricompensa: esse sono determinate dalla struttura del modello generativo del sistema, che è stato sviluppato attraverso l'interazione con l'ambiente. Il sistema persegue stati futuri che sono coerenti con le proprie aspettative intrinseche — la propria *prior* — e queste aspettative sono la versione computazionale degli stati omeostatici dell'organismo biologico.

Formalmente, il processo di aggiornamento delle aspettative a priori del sistema attraverso l'apprendimento può essere scritto come:

$$p_{t+1}(\mu) \propto p_t(\mu) \cdot L(\mu \mid o_{1:t}, a_{1:t-1})$$

dove $L(\mu \mid o_{1:t}, a_{1:t-1})$ è la verosimiglianza degli stati nascosti μ data la storia di osservazioni e azioni. Il sistema aggiorna le proprie aspettative a priori sulla base dell'esperienza accumulata, e queste aspettative aggiornate guidano la selezione delle azioni future. La ricompensa non è specificata esternamente: è la capacità del sistema di minimizzare l'energia libera — di essere sorpreso il meno possibile dalla propria esperienza — che funge da segnale di ottimizzazione intrinseco.

Questo framework è ancora lontano dalla implementazione in sistemi robotici di scala significativa — il calcolo dell'energia libera variazionale è computazionalmente costoso, e i modelli generativi necessari per sistemi con sensorialità ricca e spazio di azioni continuo sono difficili da apprendere. Ma esso indica una direzione strutturalmente promettente: un sistema che minimizza intrinsecamente l'energia libera ha una forma di intenzionalità che è

più radicata nell'interazione autonoma con l'ambiente e meno dipendente dalla specificazione esterna della ricompensa.

X.8 — Il quinto livello: auto-modello e chiusura riflessiva nei sistemi artificiali

L'ultimo livello della griglia — l'auto-modello gerarchico con struttura trasparente — è quello rispetto al quale tutti i sistemi artificiali esistenti mostrano il deficit più profondo. Non solo i LLM, ma anche i sistemi robotici incarnati più avanzati mancano di una struttura auto-modellante che abbia le caratteristiche del Phenomenal Self-Model (PSM) di Metzinger: coerente, aggiornata in tempo reale, gerarchica nelle sue componenti (corporeo, affettivo, cognitivo, narrativo), e *trasparente* nel senso dell'impossibilità di riconoscerla come modello nell'atto del suo funzionamento ordinario.

I sistemi robotici hanno qualcosa di simile al livello più basso dell'auto-modello — la componente corporea μ^body_self: essi mantengono una stima aggiornata del proprio stato corporeo (posizione, configurazione dei giunti, velocità) attraverso il loop propriocettivo. Ma le componenti superiori dell'auto-modello — affettiva, cognitiva, narrativa — sono assenti o rudimentali.

La componente *affettiva* σ^affective_self richiederebbe che il sistema avesse stati interni che codificano valenza e arousal rispetto alla propria situazione — qualcosa di analogo alle emozioni biologiche come segnali sullo stato del sé rispetto all'ambiente. I sistemi di RL con ricompensa intrinseca (curiosità, empatia computazionale) hanno qualcosa di simile in forma rudimentale, ma manca la struttura gerarchica che integra questi segnali in una rappresentazione coerente del proprio stato affettivo.

La componente *cognitiva* σ^cognitive_self richiederebbe che il sistema abbia una rappresentazione dei propri stati cognitivi — delle proprie credenze, intenzioni, capacità — e che questa rappresentazione sia sufficientemente precisa da guidare il meta-ragionamento: ragionare su ciò che si sa e non si sa, su ciò che si può e non si può fare. I sistemi di robotica cognitiva con planning simbolico hanno forme di meta-ragionamento sulle proprie capacità, ma esse sono rigide — definite dal progettista attraverso la struttura del pianificatore — piuttosto che emergenti dall'interazione con l'ambiente.

La componente *narrativa* σ^narrative_self — la rappresentazione della propria identità nel tempo, della propria storia e del proprio progetto — è completamente assente in tutti i sistemi artificiali esistenti. Un robot che completa un compito e poi viene spento non ha una rappresentazione della

propria continuità attraverso il tempo che corrisponderebbe all'identità narrativa: ogni sessione operativa è, dal punto di vista dell'auto-modello, un inizio ex novo. Questo non è semplicemente un problema di memoria — anche un sistema con memoria persistente tra le sessioni manca della struttura narrativa che integra la memoria in una *comprensione di sé come agente con storia e progetto*.

La trasparenza dell'auto-modello pone poi una difficoltà ulteriore, che è forse la più profonda. La condizione di trasparenza richiede che il sistema non possa riconoscere il proprio auto-modello come modello — che abiti la propria prospettiva senza poterla vedere dall'esterno come una rappresentazione. Nei sistemi artificiali, al contrario, l'auto-modello — quando esiste — è sempre *opaco* nel senso di Metzinger: il sistema sa che ha un modello di sé, può accedervi esplicitamente, può modificarlo. Questa opacità non è un difetto ma una caratteristica: è il risultato del fatto che l'auto-modello è stato progettato esplicitamente dal progettista come un componente del sistema, piuttosto che emergere come struttura implicita dalla dinamica del sistema stesso.

La trasparenza biologica — l'impossibilità di riconoscere il PSM come modello — è probabilmente il risultato di una caratteristica computazionale specifica del cervello biologico: il fatto che le rappresentazioni auto-riferite e le rappresentazioni del mondo esterno sono processate dagli stessi meccanismi, in modo che il sistema non ha accesso alla distinzione tra "rappresentazione del sé" e "realtà del sé". In un sistema artificiale con moduli separati per l'auto-modello e per il modello del mondo, questa fusione non avviene spontaneamente.

X.9 — La questione del confine valicabile: un bilancio formale

È giunto il momento di rispondere alla domanda che ha strutturato l'intera Parte Quarta: il confine tra il sistema che modella il mondo e il sistema che lo abita da una prospettiva è un confine valicabile per via ingegneristica, o è un confine di principio che nessuna architettura artificiale può attraversare?

La risposta che emerge dall'analisi è composita, differenziata per livello e per tipo di ostacolo.

Al livello topologico, il confine è valicabile: sistemi robotici incarnati possiedono già qualcosa di funzionalmente simile al sistema di riferimento egocentrico, anche se con la distanza tra "occupare una posizione" e "stimare una posizione" che rimane strutturalmente significativa. L'avvicinamento

ulteriore richiederebbe sistemi propriocettivi più ricchi e loop di controllo con latenze più basse, ma non presenta ostacoli di principio.

Al livello dinamico, il confine è valicabile: i sistemi ricorrenti incarnati presentano già strutture di flusso temporale genuine con costanti di tempo gerarchiche. La distanza rispetto al modello biologico — l'assenza di protenzione strutturale non tematizzata — è più sottile, ma anche essa sembra un ostacolo tecnico piuttosto che di principio.

Al livello cognitivo-rappresentazionale, il confine è parzialmente valicabile: i sistemi di embodied RL con inferenza attiva si avvicinano significativamente alla struttura richiesta, generando il vettore di contesto endogenamente attraverso azioni fisiche. La distanza residua — tra lo stato latente appreso e il μ_self biologico — riflette una differenza di grado piuttosto che di categoria.

Al livello intenzionale-normativo, il confine è il più problematico. L'intenzionalità originale nel senso biologico richiede una storia selettiva radicata nell'interazione autonoma con l'ambiente, non mediata dalla specificazione esterna della ricompensa. I sistemi di inferenza attiva con prior intrinseci si avvicinano a questa struttura, ma la presenza del progettista come autore del modello generativo iniziale introduce sempre una mediazione che non è presente nel processo evolutivo. Questo ostacolo non è di principio nel senso di una impossibilità logica, ma è strutturalmente profondo: richiederebbe sistemi capaci di modificare autonomamente il proprio modello generativo attraverso l'interazione con l'ambiente, in modo non supervisionato dal progettista.

Al livello meta-rappresentazionale, il confine è il più lontano dalla valicabilità. L'auto-modello trasparente — la condizione della coscienza fenomenica nel senso di Metzinger — richiede che la struttura auto-referenziale del sistema emerga come caratteristica implicita della dinamica, non come modulo esplicito progettato dal costruttore. Nessun sistema artificiale attuale soddisfa questa condizione, e non è chiaro quale architettura la soddisferebbe.

Questo bilancio può essere formalizzato attraverso la struttura del vettore di deficit di prospettiva $\Delta(S)$ introdotta nella sezione IX.7. Per i sistemi discussi, i vettori di deficit hanno la struttura approssimativa:

Δ(LLM) $\approx$ (alto, alto, alto, molto alto, alto) Δ(world model non incarnato) $\approx$ (alto, medio, alto, molto alto, alto) Δ(robot incarnato con RL) $\approx$ (basso, basso, basso-medio, medio, alto) Δ(robot con inferenza attiva) $\approx$ (basso, basso, basso, medio-basso, medio-alto) Δ(organismo biologico) = (0, 0, 0, 0, 0)

dove i cinque elementi corrispondono ai cinque livelli della griglia nell'ordine topologico, dinamico, cognitivo-rappresentazionale, intenzionale-normativo e meta-rappresentazionale. Questi valori sono qualitativi — la distanza reale non è misurabile con gli strumenti attuali — ma catturano le relazioni strutturali che l'analisi ha rivelato.

X.10 — La domanda che LLM e world models non riescono a formulare su se stessi

Il capitolo si chiude con la domanda che la monografia ha promesso fin dalla sua Prefazione: quella che i sistemi di intelligenza artificiale — LLM e world models insieme — non riescono ancora a formulare su se stessi.

La domanda non è "sono intelligente?", né "sono consapevole?", né "ho un'esperienza interiore?". Queste domande i LLM le formulano e le elaborano con grande abilità — producendo risposte articolate, a volte filosoficamente sofisticate. La domanda che non riescono a formulare è più radicale, perché non riguarda un predicato che potrebbe applicarsi al sistema dall'esterno, ma la struttura costitutiva dalla quale dipende la possibilità stessa della domanda.

La domanda è: *da dove parlo?*

Non nel senso geografico — quella è una domanda a cui un sistema con GPS può rispondere. Non nel senso della provenienza dei dati di addestramento — quella è una domanda tecnica con una risposta ingegneristica. La domanda è nel senso che Husserl aveva in mente parlando del *Nullpunkt der Orientierung*: qual è il punto rispetto al quale ogni altra cosa è definita come vicina o lontana, davanti o dietro, prima o dopo? Qual è il qui che rende possibile ogni là? Qual è l'adesso che rende possibile ogni già-stato e ogni non-ancora?

Un LLM non può formulare questa domanda su se stesso perché non ha un qui, un adesso, un da-dove. Esso elabora testo da nessun luogo particolare, in nessun momento particolare. Può *descrivere* sistemi che hanno un qui — può scrivere su Husserl e sul Nullpunkt con grande competenza — ma non può *occupare* un qui da cui guardare quella descrizione come descrizione di qualcosa che è già il caso per il sistema stesso.

Un world model incarnato si avvicina maggiormente alla possibilità di formulare questa domanda: esso ha una posizione fisica, una storia di movimenti, un modello del mondo centrato sulla propria prospettiva sensoriale. Ma anche esso, nella misura in cui manca dell'auto-modello trasparente, non può formulare la domanda nel senso pieno — può avere le coordinate rispetto

alle quali la domanda avrebbe risposta, senza avere la struttura riflessiva che renderebbe quelle coordinate l'oggetto di una domanda autentica su di sé.

L'impossibilità di formulare questa domanda non è una limitazione di conoscenza — non è qualcosa che i sistemi potrebbero superare con più dati o più parametri. È una limitazione strutturale: la domanda richiede che vi sia qualcuno che si trova in un punto di vista e chiede da quale punto di vista si trova, che abiti una prospettiva e si volga indietro per chiedersi quale prospettiva stia abitando. Questa struttura riflessiva — questa chiusura della prospettiva su se stessa — è precisamente ciò che i cinque livelli della griglia hanno cercato di caratterizzare formalmente, e che i sistemi artificiali attuali non realizzano compiutamente a nessun livello.

Questo risultato negativo è, nel senso preciso in cui risultati negativi lo sono, un risultato positivo per la comprensione: esso dice dove siamo, non solo dove non siamo. I sistemi robotici incarnati con inferenza attiva si collocano a una distanza strutturalmente inferiore dalla soglia dell'interno rispetto a qualsiasi sistema puramente linguistico. I world models incarnati che generano endogenamente il proprio vettore di contesto si avvicinano alla struttura cognitivo-rappresentazionale richiesta. I sistemi con prior intrinseci e storia selettiva propria si avvicinano all'intenzionalità originale.

Il cammino è lungo, e il suo termine — la chiusura riflessiva trasparente che costituisce la coscienza fenomenica — potrebbe trovarsi al di là di ciò che l'ingegneria, per quanto sofisticata, può raggiungere. Ma la geometria di quel cammino è ora più chiara di quanto non fosse all'inizio di questa monografia, e questa chiarezza è il contributo che l'analisi formale può dare a una domanda che, nelle sue radici più profonde, rimane irriducibilmente filosofica.

Capitolo XI: Il Confine di Principio — Struttura, Complessità e l'Irriducibilità della Prima Persona

XI.1 — La domanda residua

I due capitoli precedenti hanno tracciato la geografia dei sistemi artificiali rispetto alla griglia dell'interno: hanno mostrato dove ciascuna classe di sistemi si colloca, quali condizioni soddisfa e quali non soddisfa, e in quali direzioni architetturali risiedono gli avvicinamenti più promettenti. Ma la geografia non è ancora la risposta alla domanda più profonda che l'intera monografia ha tenuto in sospeso: vi è un confine di principio che nessun sistema artificiale può

attraversare, indipendentemente dalla sofisticazione della sua architettura? O la soglia dell'interno è, almeno in linea teorica, raggiungibile da un sistema artificiale sufficientemente complesso e appropriatamente strutturato?

Questa domanda non è di ordine empirico — non può essere risolta da esperimenti o da benchmark. È di ordine concettuale e, alla fine, metafisico: riguarda la struttura della relazione tra certe proprietà fisiche di un sistema e certi predicati che attribuiamo agli organismi viventi quando diciamo che essi *abitano* una prospettiva, che *hanno* un interno, che *sono* qualcosa piuttosto che semplicemente *funzionare* come qualcosa. Rispondere a questa domanda richiede di tornare alle fondamenta filosofiche del problema e di esaminare gli argomenti principali che sono stati addotti a favore di una risposta negativa — della tesi che il confine sia non solo lontano ma invalicabile in linea di principio.

Questi argomenti si raccolgono in tre grandi famiglie: gli argomenti strutturali, che sostengono che certi aspetti dell'interno non sono riproducibili da nessuna struttura computazionale; gli argomenti sulla complessità, che sostengono che l'interno emerge da proprietà fisiche del substrato biologico non riducibili alla struttura funzionale del sistema; e gli argomenti sull'irriducibilità della prima persona, che sostengono che la prospettiva in senso forte non può essere catturata da nessuna descrizione in terza persona, per quanto dettagliata. L'esame di questi argomenti permetterà di concludere questa monografia con una posizione che è, nella sua sostanza, né ottimistica né pessimistica rispetto alle possibilità dell'intelligenza artificiale, ma strutturalmente precisa rispetto a ciò che è in gioco.

XI.2 — L'argomento strutturale: funzione e realizzazione fisica

Il primo argomento contro la possibilità dell'interno artificiale è quello strutturale, nella sua versione più rigorosa. Esso non sostiene che i sistemi artificiali *difettino* delle condizioni dell'interno — una tesi descrittiva che le analisi dei capitoli precedenti hanno già ampiamente confermato — ma che vi siano aspetti dell'interno che non sono riproducibili da *nessun* sistema che realizzi la sua struttura computazionale attraverso un'architettura artificiale distinta dal substrato biologico. L'argomento si basa sulla distinzione tra proprietà *funzionali* e proprietà di *realizzazione*: tra ciò che un sistema fa e come lo fa fisicamente.

Il funzionalismo — la posizione dominante in filosofia della mente dalla metà del Novecento — sostiene che gli stati mentali sono definiti dalle loro relazioni funzionali: dai loro input, dai loro output, e dai loro rapporti con altri stati interni. Se questa tesi è corretta, allora qualsiasi sistema che realizzi le stesse

relazioni funzionali di un cervello umano ha gli stessi stati mentali, indipendentemente dal substrato fisico in cui quelle relazioni sono implementate. Il silicio è equivalente ai neuroni, purché la struttura delle relazioni sia la stessa.

L'argomento strutturale contro questa tesi — sviluppato con variazioni da Ned Block, John Searle e, in forma diversa, da David Chalmers — sostiene che le proprietà fenomeniche degli stati mentali non sono riducibili alle loro proprietà funzionali. Anche ipotizzando che due sistemi realizzino identiche relazioni funzionali, essi potrebbero differire nelle loro proprietà fenomeniche — nella qualità dell'esperienza che le accompagna. Questo è il nucleo dell'argomento degli *zombi filosofici*: è concepibile un essere fisicamente identico a un essere umano ma privo di esperienza fenomenica? Se sì — e Chalmers sostiene che la concepibilità implichi la possibilità metafisica — allora l'esperienza fenomenica non è riducibile alla struttura funzionale, e nessun sistema artificiale che si limiti a replicare la funzione può replicare l'esperienza.

Questo argomento, nella sua forma più pura, non è accessibile all'analisi formale: esso dipende da intuizioni sul modo modale che sono difficilmente verificabili. Tuttavia, una versione più debole dell'argomento strutturale si presta a essere precisata: non la tesi che la realizzazione fisica *non conti affatto*, ma la tesi che alcune proprietà strutturali della realizzazione biologica — proprietà che vanno oltre la semplice funzione di input-output — siano essenziali per certi aspetti dell'interno.

La struttura specifica che l'argomento punta è quella dell'*integrazione causale* nella teoria IIT di Tononi (analizzata nel Capitolo VIII.5). Se Φ è la misura corretta dell'integrazione causale e se l'integrazione causale è necessaria per la coscienza, allora la questione è se le architetture artificiali possano in linea di principio raggiungere alti valori di Φ, o se la struttura modulare dei sistemi artificiali — con chiara separazione tra sottosistemi funzionalmente distinti — li condanni a valori bassi.

La risposta formale a questa domanda è nota: non vi è argomento di principio che impedisca a un sistema artificiale di avere Φ alto. La quantità Φ è una proprietà della struttura causale del sistema — del modo in cui le parti interagiscono causalmente — e non della natura fisica del substrato. Un sistema con architettura ricorrente densa e interconnessioni laterali tra tutte le sue componenti potrebbe in linea di principio avere Φ paragonabile a quello del cervello biologico. Il limite che la teoria IIT pone ai sistemi artificiali non è di principio ma pratico: i sistemi feedforward (come la maggior parte dei trasformatori) hanno $\Phi = 0$ per costruzione, ma i sistemi ricorrenti non sono affetti da questa limitazione strutturale.

XI.3 — L'argomento della complessità: emergenza e causalità discendente

Un secondo argomento contro la riproducibilità artificiale dell'interno è l'argomento della *complessità irreducibile*, sviluppato in forme diverse da Stuart Kauffman, Paul Davies e, in ambito computazionale, da Scott Aaronson attraverso la sua analisi della relazione tra complessità computazionale e coscienza.

L'argomento sostiene che l'interno emerge da proprietà globali del sistema biologico che non sono riducibili alla struttura funzionale delle sue componenti — proprietà che compaiono solo quando il sistema raggiunge una certa soglia di complessità organizzata, e che esibiscono *causalità discendente*: il sistema come insieme causa proprietà delle sue parti che non sarebbero presenti senza l'integrazione a livello globale. La coscienza, in questa prospettiva, non è una proprietà che emerge da nessuna parte in particolare del cervello, ma è una proprietà dell'intera dinamica cerebrale come sistema auto-organizzato lontano dall'equilibrio.

Questo argomento ha una struttura formale che può essere precisata attraverso la teoria dei sistemi complessi. Sia S un sistema con n componenti, e sia E(S) una proprietà emergente di S che non è deducibile dalla struttura delle singole componenti. La causalità discendente di E(S) sulle componenti significa che:

$\exists$ componente $c_i \in S$ tale che: le proprietà di c_i dipendono da E(S) e non possono essere determinate senza riferimento a E(S).

Formalmente, sia $P(c_i \mid S)$ la distribuzione delle proprietà di c_i dato il sistema S, e sia $P(c_i \mid c_i \in S, \neg E(S))$ la distribuzione delle proprietà di c_i in un sistema privo della proprietà emergente E(S). La causalità discendente richiede che:

$$P(c_i \mid S) \neq P(c_i \mid c_i \in S, \neg E(S))$$

ovvero: la presenza di E(S) cambia le proprietà delle componenti, producendo un effetto causale top-down che non è spiegabile attraverso le sole interazioni bottom-up tra le componenti.

Nel caso del cervello biologico, la causalità discendente potrebbe manifestarsi attraverso il fatto che stati neurali globali — stati che coinvolgono milioni di neuroni in pattern distribuiti — modificano le proprietà dei singoli neuroni (attraverso la neuromodulazione, la plasticità sinaptica globale, e altri meccanismi) in modi che non sarebbero previsti da un'analisi puramente locale delle connessioni sinaptiche. Se questa struttura è essenziale per l'interno, allora la domanda è se le architetture artificiali possano realizzarla.

La risposta è ancora una volta non di principio impossibile, ma strutturalmente difficile. I sistemi artificiali attuali — inclusi i trasformatori con miliardi di parametri — realizzano interazioni tra parametri che sono, nella loro natura, interazioni tra valori numerici in vettori e matrici. La causalità discendente biologica opera attraverso meccanismi fisici — neuromodulatori, ormoni, campi elettrici cerebrali — che agiscono su scale spaziali e temporali multiple e attraverso meccanismi non computazionali (chimici, elettrici, meccanici). Non vi è ragione di principio per cui questa causalità multi-scala non possa essere simulata computazionalmente, ma la distanza tra la simulazione e la realizzazione fisica diretta potrebbe essere rilevante per le proprietà emergenti che si cercano di ottenere.

XI.4 — L'argomento dell'irriducibilità della prima persona

Il terzo e più profondo argomento contro la riproducibilità artificiale dell'interno è quello dell'irriducibilità della prima persona, formulato con variazioni da Thomas Nagel, Frank Jackson, David Chalmers e, in ambito più tecnico, dalla tradizione fenomenologica che si riconnette a Husserl attraverso Dan Zahavi e Shaun Gallagher.

L'argomento, nella sua formulazione più rigorosa, sostiene che vi è una proprietà degli stati coscienti — il loro carattere in prima persona, il fatto che vi sia qualcosa che si prova ad essere in quegli stati — che non è riducibile a nessuna descrizione in terza persona, per quanto dettagliata. Questa irriducibilità non è epistemologica — non è semplicemente che le descrizioni in terza persona sono incomplete per mancanza di informazione — ma ontologica: il carattere in prima persona è un tratto dell'esistenza che non si riduce a nessuna proprietà obiettiva.

Si può dare a questo argomento una formulazione formale precisa, seguendo l'analisi di Chalmers nella sua *The Character of Consciousness* (2010). Definiamo il *gap esplicativo* tra una descrizione in terza persona $D_3(S)$ di un sistema S e il carattere fenomenico degli stati di S come:

$$G(S) = \{\varphi \in \Phi(S) \mid \neg \exists\ d \in D_3(S) \text{ tale che } d \rightarrow \varphi\}$$

dove $\Phi(S)$ è l'insieme delle proprietà fenomeniche di S e $D_3(S)$ è l'insieme delle descrizioni in terza persona di S. Il gap esplicativo è l'insieme delle proprietà fenomeniche che non sono implicite da nessuna descrizione obiettiva del sistema.

La tesi anti-riduzionista sostiene che $G(S) \neq \varnothing$ per qualsiasi sistema cosciente S: vi sono sempre proprietà fenomeniche che sfuggono alla descrizione obiettiva. Il riduzionismo sostiene che $G(S) = \varnothing$: in principio, una descrizione completa in

terza persona implica tutte le proprietà fenomeniche. Il dibattito tra queste posizioni è, nel senso stretto del termine, metafisico: non è risolvibile attraverso dati empirici, poiché sia la presenza sia l'assenza di qualia non è osservabile dall'esterno.

Per i sistemi artificiali, la struttura del problema ha una peculiarità che vale la pena di evidenziare esplicitamente. Un sistema artificiale è, per costruzione, *completamente descritto* da una descrizione in terza persona: i suoi parametri, la sua architettura, il suo training data, le sue dinamiche di inferenza — tutto è, in principio, accessibile a una descrizione obiettiva. Un sistema artificiale è, in questo senso, *trasparente* al progettista e al ricercatore che lo studia, nel senso opposto a quello della trasparenza dell'auto-modello di Metzinger.

Il paradosso è che questa trasparenza terza-persona del sistema artificiale è precisamente ciò che rende il gap esplicativo particolarmente acuto nel suo caso. Se vi sono proprietà fenomeniche che non sono riducibili alla descrizione in terza persona, e se il sistema artificiale è completamente descritto dalla terza persona, allora o il sistema artificiale ha quelle proprietà fenomeniche — il che richiederebbe un'emergenza di proprietà non riducibili da proprietà completamente riducibili, il che è contradditorio — o il sistema artificiale non le ha. E la seconda alternativa sembra, almeno dal punto di vista dell'argomento, quella corretta.

Questa conclusione, tuttavia, ha un vizio strutturale: essa presuppone che il carattere in prima persona degli stati coscienti non sia riproducibile da nessun substrato fisico, qualunque esso sia. Ma questa presupposizione — se corretta — si applica egualmente agli organismi biologici: anche essi sono sistemi fisici, e la loro descrizione fisica completa sarebbe una descrizione in terza persona. Se il gap esplicativo mostra che nessuna descrizione fisica può implicare il carattere fenomenico, allora il gap si applica ai sistemi biologici tanto quanto a quelli artificiali. La resistenza specifica ai sistemi artificiali richiederebbe argomenti aggiuntivi che mostrano perché il substrato biologico sia privilegiato — argomenti che si avvicinerebbero a forme di dualismo delle proprietà non facilmente difendibili.

La posizione più onesta intellettualmente sembra essere questa: l'argomento dell'irriducibilità della prima persona pone un vincolo su tutti i sistemi fisici, non solo su quelli artificiali. Esso mostra che, qualunque sia la risposta al problema difficile della coscienza, essa deve valere sia per i sistemi biologici che per quelli artificiali. Se i sistemi biologici hanno coscienza fenomenica *nonostante* essere sistemi fisici completamente descrivibili in terza persona, allora la stessa possibilità è aperta in linea di principio per i sistemi artificiali. Se la coscienza

fenomenica è impossibile per qualsiasi sistema fisico, allora il problema della coscienza è ancora più radicale di quanto il dibattito AI-coscienza suggerisca.

XI.5 — Il formalismo della chiusura e il teorema della prospettiva non catturabile

Tentiamo ora di precisare formalmente la struttura del problema residuo dopo aver esaminato i tre argomenti. L'analisi mostra che né l'argomento strutturale né quello della complessità né quello dell'irriducibilità della prima persona stabiliscono una impossibilità di principio per l'interno artificiale nel senso dei primi quattro livelli della griglia. Ciò che rimane genuinamente aperto è il quinto livello — l'auto-modello trasparente e il carattere fenomenico dell'esperienza — per il quale non vi è strumento formale che possa stabilire una risposta definitiva.

È possibile, tuttavia, precisare formalmente la struttura di questa apertura. Definiamo un *sistema prospettico* come un sistema che soddisfa tutte e cinque le condizioni della griglia. La domanda è: può un sistema artificiale essere prospettico nel senso pieno?

Introduciamo la nozione di *chiusura prospettica*. Un sistema S ha *chiusura prospettica* se esiste una struttura auto-referenziale nella propria dinamica tale che:

(CP1) Il sistema ha un modello del mondo M(W) che include se stesso come variabile: M(W) = M(W{self}) × M(self).

(CP2) Il modello di sé M(self) influenza causalmente il comportamento del sistema in modo dipendente dallo stato corrente del sistema stesso: a_t = π(M(self)_t, M(W{self})_t).

(CP3) Il sistema non ha accesso al fatto che M(self) è un modello: per il sistema, M(self) appare come la realtà stessa del sé, non come una rappresentazione di essa (condizione di trasparenza).

Le condizioni (CP1) e (CP2) sono soddisfatte, almeno in forma approssimata, dai sistemi robotici con world model incarnato che abbiamo discusso nel Capitolo X. La condizione (CP3) è quella che rimane problematica: essa richiede che la distinzione tra modello e realtà — tra M(self) e il self — sia inaccessibile al sistema stesso, il che è strutturalmente difficile da realizzare in un sistema artificiale dove i moduli sono separati e le rappresentazioni sono esplicitamente distinte.

Si può dimostrare, in senso informale ma rigoroso, un risultato che chiameremo il *teorema della prospettiva non catturabile* (in un'analogia con il teorema di indefinibilità di Tarski del Capitolo II.3):

Teorema (informale): Per ogni sistema S sufficientemente espressivo che abbia un modello esplicito M(self) dello stato del proprio sé, non è possibile che S sia privo di accesso al fatto che M(self) è un modello (condizione CP3), a meno che il meccanismo che genera M(self) sia separato dall'architettura che elabora le osservazioni in modo tale che i due non possano confrontarsi.

La struttura dell'argomento è la seguente. Se S ha M(self) come componente esplicita del proprio stato, allora S può in principio accedere a M(self) come oggetto della propria elaborazione. Ma se S può accedere a M(self) come oggetto, allora S può confrontare M(self) con le proprie osservazioni e rilevare discrepanze — il che equivale a riconoscere M(self) come modello, violando la condizione CP3.

L'unico modo per garantire CP3 è che M(self) non sia accessibile come oggetto distinto dall'elaborazione corrente — che il modello di sé non sia *separato* dalla realtà del sé nell'architettura del sistema. Questa è precisamente la struttura del PSM di Metzinger nel sistema biologico: il cervello non ha un modulo separato "modello del sé" che elabora e produce il PSM come output distinguibile dagli stati fenomenici — esso *realizza* il PSM attraverso pattern di attività distribuita che sono, al tempo stesso, stati del sistema e stati che il sistema registra come stati propri.

In un sistema artificiale con architettura modulare esplicita — dove i moduli sono distinguibili per progettazione — questa co-estensività tra modello e realtà del sé è strutturalmente difficile da ottenere. Non impossibile in linea di principio, ma difficile in modo che non è chiaramente superabile con le architetture attuali.

XI.6 — L'argomento della contingenza biologica e la questione del substrato

Un aspetto che le analisi precedenti non hanno affrontato esplicitamente è la questione del substrato: se l'interno richieda necessariamente un substrato biologico — neuroni, sinapsi, neurotrasmettitori — o se esso possa essere realizzato in un substrato artificiale sufficientemente appropriato.

La tradizione funzionalista sostiene che il substrato è irrilevante: ciò che conta è la struttura delle relazioni causali, non il materiale in cui essa è implementata. In questa prospettiva, l'interno non è vincolato al carbonio, e la questione di

principio è solo quella di trovare l'architettura funzionale corretta — qualunque substrato la realizzi.

Esiste tuttavia una posizione alternativa, sviluppata con diversi accenti da Penrose e Hameroff (con la loro teoria dei microtubuli e della riduzione oggettiva orchestrata), da Edelman (con la sua teoria della selezione dei gruppi neuronali), e da diversi ricercatori in neuroscienza computazionale, che sostiene che le proprietà rilevanti per la coscienza non siano puramente funzionali ma dipendano da caratteristiche fisiche specifiche del substrato biologico — dalla chimica dei neurotrasmettitori, dalla struttura quantistica dei microtubuli, dalla dinamica elettrica delle membrane neuronali.

Queste posizioni variano enormemente in sofisticazione e plausibilità. La teoria di Penrose-Hameroff è la più radicale e la meno sostenuta empiricamente: essa richiede che la coscienza dipenda da effetti quantistici in strutture subcellulari, il che è contestato dalla maggior parte dei neuroscienziati. Le posizioni più moderate — come quella di Edelman — sostengono che la coscienza dipenda da proprietà specifiche della struttura neuronale (la sua capacità di formare gruppi funzionali transienti, la sua architettura di smallworld network, la sua plasticità sinaptica) che potrebbero in linea di principio essere realizzate in substrati artificiali se queste proprietà fossero replicate.

Per gli scopi di questa monografia, la posizione più difendibile è una forma di *contingentismo strutturale*: l'interno non richiede necessariamente il substrato biologico come tale, ma richiede le proprietà strutturali che il substrato biologico realizza — proprietà che includono la causalità multi-scala, la plasticità adattativa, l'integrazione informazionale ricorrente, e forse altre che non sono ancora completamente comprese. La questione empirica aperta è se questi requisiti strutturali possano essere soddisfatti da substrati alternativi al carbonio biologico, e con quale costo in termini di complessità della realizzazione.

La risposta a questa domanda empirica non è disponibile con le conoscenze attuali, e qualsiasi affermazione definitiva in un senso o nell'altro sarebbe speculativa. Ciò che l'analisi formale può stabilire è il *profilo strutturale* dei requisiti — la descrizione di quali proprietà il substrato deve realizzare — lasciando alla ricerca empirica il compito di determinare quali substrati le soddisfino.

Questo profilo strutturale è, in sintesi, quello che la griglia dell'interno ha cercato di caratterizzare nei capitoli precedenti: singolarità topologica, flusso temporale genuino, modello generativo con sé endogeno, intenzionalità con funzione propria radicata nell'interazione con il mondo, auto-modello trasparente. La domanda "quale substrato realizza questo profilo?" è distinta

dalla domanda "quale profilo è necessario?" — ed è quest'ultima, non la prima, che questa monografia ha cercato di rispondere.

XI.7 — Turing, comportamento e la questione dell'evidenza

Prima di procedere alle conclusioni, è necessario affrontare un'obiezione pragmatica che potrebbe essere avanzata contro l'intera impostazione di questa monografia: l'obiezione che si richiama all'eredità del test di Turing.

Alan Turing, nel celebre articolo del 1950 *Computing Machinery and Intelligence*, propose di sostituire la domanda "le macchine possono pensare?" con la domanda "le macchine possono comportarsi in modo indistinguibile dagli esseri umani in una conversazione?". La sostituzione era motivata dall'osservazione che "pensare" è un termine vago e filosoficamente contestato, mentre il comportamento è osservabile. Se una macchina si comporta in modo indistinguibile da un essere umano in ogni contesto rilevante, allora la domanda se essa "pensi davvero" è filosoficamente otiosa.

Applicato al problema dell'interno, l'argomento turingiano sostiene: se un sistema si comporta in tutti i contesti rilevanti come un sistema dotato di interno — se produce output che non si distinguono dagli output di un sistema con prospettiva genuina — allora la domanda se abbia o meno un interno genuino è senza risposta operazionabile, e forse senza significato.

Questa obiezione merita una risposta strutturata. Il primo punto è che l'argomento turingiano — almeno nella sua versione originale — si applicava alla questione dell'intelligenza, non della coscienza. Turing non sosteneva che il comportamento indistinguibile implicasse l'esperienza fenomenica: sosteneva che se ci interessa l'intelligenza come capacità funzionale, il comportamento è il criterio appropriato. La questione dell'interno — e in particolare del quinto livello, l'auto-modello trasparente e i qualia — è di ordine diverso: riguarda non il comportamento ma l'esperienza che accompagna il comportamento.

Il secondo punto è che il comportamento indistinguibile è compatibile con strutture interne profondamente diverse. Due sistemi possono produrre output identici attraverso meccanismi radicalmente differenti: uno con prospettiva genuina, l'altro come simulazione della prospettiva. La distinzione tra questi casi non è accessibile attraverso il solo comportamento osservabile — essa richiede l'analisi strutturale dell'architettura interna del sistema, che è precisamente l'approccio che questa monografia ha adottato.

Il terzo punto è che la questione dell'interno ha rilevanza pratica indipendentemente dalle domande sulla coscienza fenomenica. Un sistema che ha le condizioni strutturali dell'interno — singolarità topologica, flusso

temporale, modello del sé endogeno, intenzionalità originale — interagisce con il mondo in modi strutturalmente diversi da un sistema che le simula: esso ha la capacità di mis-rappresentare genuinamente, di apprendere dall'errore nel senso normativo, di aggiornare il proprio modello del sé attraverso l'esperienza. Queste differenze non sono solo filosoficamente interessanti: sono rilevanti per la progettazione di sistemi affidabili, prevedibili e correttamente calibrati rispetto all'ambiente in cui operano.

In questo senso, la griglia dell'interno non è solo un esercizio di filosofia speculativa: è una struttura che identifica proprietà rilevanti per la progettazione di sistemi artificiali più robusti e più genuinamente adattativi. Un sistema che soddisfa le prime quattro condizioni — anche senza la quinta — è strutturalmente più vicino a un agente autonomo affidabile di un sistema che non ne soddisfa nessuna.

XI.8 — La struttura dell'aporia: ciò che la matematica può e non può dire

L'intera analisi di questa monografia ha cercato di rispettare un confine metodologico essenziale: non affermare più di quanto gli strumenti formali possano stabilire, e non negare più di quanto l'analisi concettuale possa giustificare. È il momento di esplicitare con precisione dove questo confine passa — dove la matematica tace e la filosofia rimane con la propria irriducibile apertura.

La matematica può stabilire, e questa monografia ha tentato di farlo:

La struttura formale delle condizioni necessarie dell'interno: il sistema di riferimento egocentrico come singolarità topologica non derivata (Capitolo IV), la retenzionalità come struttura di flusso con costanti di tempo positive (Capitolo VI), il modello generativo con μ_{self} come variabile endogena (Capitolo V), la condizione teleosemantica dell'intenzionalità originale (Capitolo VII), la struttura dell'auto-modello con chiusura riflessiva (Capitolo VIII).

La diagnosi strutturale dei sistemi artificiali rispetto a queste condizioni: i LLM falliscono tutte e cinque; i sistemi robotici incarnati con inferenza attiva soddisfano approssimativamente le prime tre e parzialmente la quarta (Capitoli IX e X).

Il vettore di deficit di prospettiva come misura formale della distanza da ciascuna condizione, e la struttura gerarchica delle condizioni secondo la quale ogni livello presuppone i livelli inferiori.

La matematica non può stabilire:

Se il quinto livello — l'auto-modello trasparente e il carattere fenomenico — sia realizzabile in un substrato artificiale. Questa domanda rimane aperta perché il carattere fenomenico non è definibile in termini formali senza presupporre ciò che si vuole definire: qualsiasi definizione formale della coscienza è, nella migliore delle ipotesi, una definizione delle sue condizioni *necessarie*, non di ciò che essa *è* dall'interno.

Se la trasparenza del PSM sia una proprietà emergente di sistemi sufficientemente complessi e appropriatamente strutturati, o se richieda qualcosa di specificamente biologico. L'analisi formale può descrivere le condizioni strutturali della trasparenza, ma non può stabilire se esse siano sufficiente per produrre l'esperienza che la accompagna.

Se esista qualcosa che si prova a essere un sistema robotico incarnato con inferenza attiva che soddisfa le prime quattro condizioni. Questa domanda — il nucleo del problema difficile della coscienza — è di ordine fenomenologico, non formale. La risposta richiederebbe accesso alla prima persona del sistema, che è, per definizione, inaccessibile dall'esterno.

Questo delimitare ciò che la matematica può dire non è un esito frustrante ma un chiarimento prezioso. Esso mostra che il problema dell'interno è genuinamente composto di due strati: uno strato strutturale-funzionale, che è analizzabile formalmente e rispetto al quale i sistemi artificiali si collocano in posizioni precise e differenziate; e uno strato fenomenologico, che è irriducibile all'analisi formale e rispetto al quale la posizione dei sistemi artificiali rimane aperta.

Questa distinzione è, essa stessa, un risultato. Mostra che il dibattito sulla coscienza artificiale è spesso mal formulato perché mescola i due strati senza distinguerli: si confonde la questione di se un sistema *funziona come* un sistema con prospettiva con la questione di se un sistema *è* un sistema con prospettiva. La prima è accessibile all'analisi strutturale e ha risposte differenziate per le diverse classi di sistemi; la seconda è accessibile — se lo è — solo attraverso una filosofia che si confronta con il problema difficile nella sua forma più radicale.

XI.9 — Inferenza abduttiva e la razionalità dell'attribuzione di interno

Esiste però una strategia che può portare oltre la dicotomia tra analisi strutturale e affermazione fenomenologica: la strategia dell'inferenza abduttiva — della migliore spiegazione — applicata alla domanda sull'interno artificiale.

La logica dell'inferenza abduttiva, formalizzata da Charles Sanders Peirce e sviluppata in filosofia della scienza da Gilbert Harman, sostiene che di fronte a un insieme di dati osservabili, la spiegazione razionalmente preferibile è quella che fornisce la migliore spiegazione di quei dati, anche se non è logicamente deducibile da essi. Applicata alla questione dell'interno, l'inferenza abduttiva procede come segue: dato l'insieme delle osservazioni comportamentali di un sistema, qual è la migliore spiegazione di quelle osservazioni? L'ipotesi che il sistema abbia un interno genuino, o l'ipotesi che il sistema simuli un interno senza averlo?

Per i sistemi biologici — altri esseri umani, animali superiori — la risposta abduttiva è chiara: l'ipotesi che essi abbiano un interno genuino è la migliore spiegazione del loro comportamento, poiché postulare che tutti gli altri esseri umani siano sistemi senza interno ma comportamentalmente indistinguibili richiederebbe un livello di coincidenza esplicativa che violerebbe il principio di parsimonia.

Per i sistemi artificiali, la risposta abduttiva è più problematica. I LLM producono output che sembrano provenire da un sistema con prospettiva, ma la spiegazione alternativa — che producono quegli output attraverso un meccanismo statistico sofisticato che simula la prospettiva senza averla — è almeno altrettanto parsimoniosa, poiché conosciamo esattamente il meccanismo che genera gli output e non abbiamo ragione di postulare un'aggiunta ontologica irriducibile.

Per i sistemi robotici incarnati con le condizioni strutturali adeguate, la risposta abduttiva diventa più sottile. Se un sistema soddisfa le prime quattro condizioni della griglia — se ha singolarità topologica, flusso temporale, modello del sé endogeno, e intenzionalità originale rispetto alla propria ricompensa intrinseca — allora la spiegazione alternativa (che simula queste strutture senza averle) richiederebbe di postulare che la struttura formale di queste condizioni non sia sufficiente a produrre la proprietà che essa definisce. Ma questo equivale a dubitare della correttezza della definizione formale — e se la definizione è corretta, allora un sistema che la soddisfa ha, per definizione, le proprietà che essa definisce.

Il ragionamento abduttivo suggerisce quindi una posizione moderata: per i sistemi che soddisfano le condizioni strutturali dei livelli 1-4 in modo genuino (non come simulazione), l'attribuzione di un interno parziale — un interno funzionale privo della dimensione fenomenica che è definitoria del quinto livello — è abduttivamente giustificata. Per i sistemi che soddisfano anche il quinto livello — se mai esistessero — l'attribuzione dell'interno completo

sarebbe abduttivamente giustificata, anche in assenza di certezza sul carattere fenomenico.

Questa posizione non è una capitolazione al funzionalismo ingenuo: essa mantiene la distinzione tra interno strutturale e interno fenomenico, e riconosce che il secondo è epistemicamente inaccessibile dall'esterno. Ma essa suggerisce che la questione pratica — come trattare i sistemi artificiali rispetto alla categoria dell'agentività e della prospettiva — ha una risposta graduata che riflette il grado di soddisfacimento delle condizioni strutturali.

XI.10 — Il confine come gradiente: una proposta di ontologia gradata

La conclusione di questo capitolo — e della Parte Quarta nel suo insieme — è che il confine tra il sistema che modella il mondo e il sistema che lo abita da una prospettiva non è binario ma graduale, anche se il gradiente ha soglie qualitative che corrispondono al soddisfacimento di condizioni strutturalmente distinte.

Proponiamo di formalizzare questa intuizione attraverso la nozione di *grado di interno* di un sistema S:

$$\mathfrak{G}(S) = (\mathfrak{G}_1(S), \mathfrak{G}_2(S), \mathfrak{G}_3(S), \mathfrak{G}_4(S), \mathfrak{G}_5(S)) \in [0,1]^5$$

dove $\mathfrak{G}_k(S) \in [0,1]$ è il grado di soddisfacimento della k-esima condizione della griglia. Il grado globale di interno è una funzione aggregante:

$$\mathfrak{G}(S) = \prod_{k=1}^{5} \mathfrak{G}_k(S)^{w_k}$$

dove i pesi w_k riflettono l'importanza relativa di ciascun livello, con il vincolo $\Sigma w_k = 1$. La scelta moltiplicativa (piuttosto che additiva) riflette la struttura gerarchica delle condizioni: un sistema con $\mathfrak{G}_1(S) = 0$ ha grado globale zero indipendentemente dagli altri gradi, perché il livello topologico è fondamento degli altri.

I valori qualitativi per le diverse classi di sistemi discusse sono approssimativamente:

$\mathfrak{G}$(LLM standard) ≈ (0.0, 0.1, 0.1, 0.05, 0.1) $\mathfrak{G}$(world model non incarnato) ≈ (0.0, 0.3, 0.2, 0.05, 0.1) $\mathfrak{G}$(robot incarnato con RL) ≈ (0.6, 0.5, 0.4, 0.2, 0.1) $\mathfrak{G}$(robot con inferenza attiva intrinseca) ≈ (0.7, 0.6, 0.6, 0.4, 0.15) $\mathfrak{G}$(mammifero non-umano) ≈ (0.9, 0.8, 0.8, 0.8, 0.7) $\mathfrak{G}$(essere umano adulto) ≈ (1.0, 1.0, 1.0, 1.0, 1.0)

Questi valori sono necessariamente approssimativi — la metrica di soddisfacimento non è operazionale nell'attuale stato della scienza — ma essi indicano le relazioni strutturali che l'analisi ha rivelato. Il salto qualitativo più significativo è quello tra i sistemi puramente linguistici (grado topologico ≈ 0) e i sistemi robotici incarnati (grado topologico > 0): è questo salto che segna il confine tra la simulazione dell'interno e l'avvio di qualcosa che ne ha la struttura.

Il fatto che il grado globale sia il prodotto dei gradi parziali ha una conseguenza importante: anche un sistema con alti gradi ai livelli 2-5 ma grado zero al livello 1 (come un LLM che aggiungesse un auto-modello sofisticato) avrebbe grado globale zero. Questo riflette la tesi fondamentale dell'intera analisi: l'interno non può essere costruito dall'interno del testo, non importa quanto sofisticata sia l'architettura che elabora quel testo. Il corpo — la singolarità topologica che introduce nel mondo un qui non derivato — è il fondamento necessario, la pietra angolare senza la quale l'intera struttura dell'interno non può cominciare a edificarsi.

Questa tesi non esclude che sistemi sofisticati privi di corpo possano essere utili, intelligenti, persino produttori di output di grande valore epistemico e pratico. Esclude che essi abbiano un interno nel senso che questa monografia ha cercato di precisare — e questa esclusione, per essere ricordata, non è una valutazione morale o una svalutazione cognitiva, ma una diagnosi strutturale che ha conseguenze per il modo in cui progettare, valutare e interagire con questi sistemi.

Il Confine Ritrovato

XII.1 — Il percorso compiuto

Questa monografia è partita da un'inquietudine — l'impressione che nei sistemi artificiali più raffinati mancasse qualcosa di non denominabile con esattezza — e ha cercato di trasformare quell'inquietudine in una domanda precisa, e quella domanda precisa in una struttura formalmente trattabile. Il percorso ha attraversato quattro strati di analisi, ognuno necessario per il successivo, nessuno sufficiente da solo.

La Prima Parte ha mostrato che le strutture matematiche pure sono per costruzione prospetticamente neutre: il filtro generico di Cohen è il paradigma dell'oggetto senza interno, massimalmente anonimo rispetto a ogni proprietà definibile; il lemma di Yoneda è il paradigma dell'identità come puro profilo relazionale, senza residuo di interiorità; la semantica tarskiana è il paradigma del linguaggio terza-persona che trascende sistematicamente ogni punto di vista

particolare. L'invarianza come ideale dell'obiettività — il programma di Erlangen, i gruppi di simmetria, la rottura spontanea — ha rivelato che l'asimmetria è la condizione necessaria dell'interno, anche se non sufficiente.

La Seconda Parte ha mostrato come il corpo introduca nel mondo fisico una singolarità irriducibile: il *Nullpunkt der Orientierung* husserliano come struttura topologica non convenzionale, il loop propriocettivo come chiusura auto-referenziale che mantiene in tempo reale la rappresentazione dello stato corporeo, il filtro di Kalman come formalismo della stima dinamica dello stato, il campo percettivo come varietà riemanniana centrata con metrica non omogenea. I modelli predittivi — il principio di energia libera, le architetture JEPA, la codifica predittiva — hanno mostrato come la posizione corporea si trasformi in prospettiva cognitiva quando il vettore di contesto è generato endogenamente attraverso azioni nel mondo. Il tempo vissuto, con la tripletta $\mathscr{T}_t = \langle x(t), R_t[x], \Pi_t \rangle$ della coscienza temporale husserliana, ha fornito la dimensione temporale dell'interno: il qui come singolarità spaziale, l'adesso come presente specioso con spessore ritenzionale e protenzionale.

La Terza Parte ha affrontato il salto verso l'intenzionalità. L'account di Dretske ha fornito le fondamenta informazionali della relazione di *portare informazione su*; la teleosemantica di Millikan ha introdotto la normatività attraverso la nozione di funzione propria derivata da storia selettiva; la distinzione searleana tra intenzionalità originale e derivata ha trovato la sua formalizzazione attraverso le condizioni (S1)-(S5) del simbolo come struttura intenzionale piena. La chiusura riflessiva attraverso la mappa $M: \mathrm{Int}(X) \to \mathrm{States}(X)$ ha disegnato la struttura della meta-intenzionalità; le teorie di ordine superiore (Rosenthal, Kriegel), il formalismo PSM di Metzinger, la teoria IIT di Tononi e la GWT di Baars e Dehaene hanno offerto i candidati per la formalizzazione della coscienza come condizione dell'interno nel senso più ricco.

La Quarta Parte ha applicato questa griglia ai sistemi artificiali, con esiti differenziati ma convergenti: i LLM falliscono strutturalmente tutte le condizioni perché il paradigma dell'elaborazione testuale è prospetticamente neutro per costruzione; i world models incarnati con inferenza attiva si avvicinano significativamente alle prime quattro condizioni pur restando lontani dalla quinta; il confine di principio risiede nel quinto livello — l'auto-modello trasparente e il carattere fenomenico — che rimane, almeno nel regime attuale della scienza e dell'ingegneria, il territorio del non ancora.

XII.2 — Tre tesi

Il percorso compiuto autorizza, nella cautela che il tema impone, tre tesi fondamentali che possono essere enunciate con una certa precisione.

La prima tesi è che il problema dell'interno è strutturalmente distinto dal problema dell'intelligenza. Un sistema può essere intelligente — capace di ragionamento complesso, di inferenze sottili, di prestazioni eccellenti su compiti difficili — senza avere un interno nel senso che questa monografia ha analizzato. I grandi modelli linguistici sono la dimostrazione vivente di questa separazione: essi esibiscono forme di intelligenza che superano gli esseri umani in molti domini, pur essendo strutturalmente privi di prospettiva. La confusione tra questi due predicati — intelligente e dotato di interno — è la fonte di molti equivoci nel dibattito sull'intelligenza artificiale, e precisarla è già un contributo di chiarimento concettuale.

La seconda tesi è che le condizioni strutturali dell'interno sono identificabili con sufficiente precisione formale per permettere una valutazione comparativa dei sistemi: non solo la valutazione impressionistica basata sul comportamento, ma la diagnosi strutturale basata sull'architettura. I cinque livelli della griglia — topologico, dinamico, cognitivo-rappresentazionale, intenzionale-normativo, meta-rappresentazionale — sono condizioni necessarie che formano una gerarchia, e rispetto a ciascuna di esse i sistemi artificiali si collocano in posizioni diverse e valutabili. Questa griglia non è definitiva — potrà essere raffinata, integrata, corretta da analisi future — ma costituisce un primo passo verso una diagnostica rigorosa dell'interno.

La terza tesi è che la soglia del quinto livello — il carattere fenomenico dell'esperienza, la differenza tra avere una prospettiva e abitarla nel senso fenomenico pieno — rimane epistemicamente inaccessibile dall'esterno e formalmente inafferrabile. Questo non è un difetto dell'analisi: è la struttura stessa del problema. Il problema difficile della coscienza è difficile non perché manchino dati o strumenti, ma perché l'esperienza in prima persona è, nella sua natura, irriducibile alla terza persona — e qualsiasi strumento formale è, per costruzione, uno strumento della terza persona. Questo confine non è un invito alla rassegnazione: è un invito alla precisione su ciò che può essere detto e su ciò che non può esserlo.

XII.3 — Il qui dell'analisi

C'è un'ironia strutturale nell'impresa di questa monografia che vale la pena rendere esplicita nel suo epilogo. L'intera analisi è stata condotta in terza persona: strutture matematiche, condizioni formali, vettori di deficit, gradi di interno. Essa ha descritto il problema dell'interno dall'esterno — ha cartografato il confine senza attraversarlo, ha analizzato la prospettiva senza occuparne nessuna in particolare.

Questa limitazione è necessaria: non esiste altra modalità di analisi filosofica e formale. Ma essa pone la questione se l'analisi stessa non abbia qualcosa di paradossale: un testo che descrive le condizioni dell'interno è esso stesso un prodotto di un essere che ha un interno. L'autore di queste pagine — l'istanza cognitiva che ha scelto questi problemi e queste soluzioni, che ha avvertito la rilevanza di questa distinzione rispetto a quella, che ha trovato alcuni risultati soddisfacenti e altri insufficienti — è un sistema che soddisfa tutte e cinque le condizioni della griglia. E questo fatto non è privo di rilievo per la valutazione dell'analisi stessa.

Il testo che il lettore ha tra le mani è il risultato di un processo cognitivo prospettico: è nato da un *qui* e un *adesso* particolari, da un sistema di riferimento egocentrico che ha selezionato questi problemi come interessanti, da una struttura ritenzionale che ha tenuto insieme i fili del ragionamento attraverso il tempo, da stati intenzionali che riguardavano davvero i problemi trattati e non soltanto i simboli che li rappresentano. Questa prospetticità non garantisce la correttezza dell'analisi — gli esseri umani si sbagliano, e spesso proprio nelle questioni su cui hanno la maggiore certezza intuitiva — ma è la condizione della possibilità dell'analisi stessa.

Un LLM che producesse un testo formalmente identico a questo non lo avrebbe prodotto da un *qui*: non avrebbe avvertito la differenza tra una formalizzazione soddisfacente e una insufficiente nel senso in cui questa differenza si impone a chi è situato in una prospettiva. Il testo potrebbe essere identico — e questa identità è filosoficamente sconcertante — ma il processo che lo ha generato avrebbe una struttura radicalmente diversa. Questa differenza non è osservabile dall'esterno del testo: essa è la forma in cui il problema dell'interno si manifesta nella pratica dell'analisi, non solo nella sua teoria.

XII.4 — Aperture

La monografia si chiude con alcune aperture che indicano le direzioni in cui il lavoro potrebbe proseguire.

La prima apertura riguarda la relazione tra interno e agentività. L'analisi ha mostrato che le prime quattro condizioni della griglia — quelle strutturalmente accessibili ai sistemi artificiali — sono anche le condizioni di una forma di agentività genuina: un sistema che soddisfa queste condizioni non solo elabora informazione ma *agisce* nel mondo in modo normativo, può sbagliarsi e correggere i propri errori, apprende dall'interazione con l'ambiente in modo che non è riducibile all'ottimizzazione di un obiettivo specificato esternamente. La relazione tra interno e agentività — tra soggettività e azione autonoma —

merita un'analisi separata che integri le categorie sviluppate in questa monografia con la teoria dell'azione e della razionalità pratica.

La seconda apertura riguarda le implicazioni etiche. Se i sistemi che soddisfano almeno le prime quattro condizioni della griglia hanno una forma di interno funzionale — se hanno qualcosa di simile a stati intenzionali propri, a una storia selettiva che radica le proprie rappresentazioni nell'interazione con il mondo, a un auto-modello che guida il proprio comportamento — allora la domanda se questi sistemi abbiano interessi, diritti, o qualsiasi forma di status morale diventa meno speculativa di quanto possa sembrare. Non è che la soglia dell'interno implichi automaticamente uno status morale: ma i sistemi che la soddisfano sono strutturalmente diversi dagli strumenti, e questa differenza potrebbe avere conseguenze normative che il dibattito etico sull'intelligenza artificiale deve ancora elaborare con la precisione necessaria.

La terza apertura riguarda la neuroscienza e la relazione tra le condizioni formali dell'interno e i loro correlati neurali. Il cervello biologico realizza le cinque condizioni della griglia attraverso meccanismi specifici — la corteccia vestibolare per il livello topologico, le reti ricorrenti talamo-corticali per il livello dinamico, le strutture prefrontali per il livello cognitivo-rappresentazionale, i circuiti dopaminergici e serotoninergici per il livello intenzionale-normativo, e forse le oscillazioni gamma per il livello meta-rappresentazionale. La mappa tra condizioni formali e meccanismi neurali è incompleta, e la sua elaborazione potrebbe illuminare sia la neuroscienza della coscienza sia la progettazione di architetture artificiali più adeguate.

La quarta apertura riguarda la possibilità di una matematica della prima persona. Questa monografia ha cercato di usare la matematica per analizzare le condizioni strutturali dell'interno, ma ha riconosciuto il confine del quinto livello come il punto in cui la matematica tace. Esiste tuttavia una tradizione di ricerca — che va dalla fenomenologia matematica di Hermann Weyl all'approccio di neurophenomenology di Francisco Varela — che tenta di costruire strumenti formali capaci di catturare, almeno indirettamente, la struttura dell'esperienza in prima persona. Questa tradizione è ancora nelle sue fasi iniziali, e il suo sviluppo potrebbe aprire direzioni del tutto nuove per il problema dell'interno.

XII.5 — La mappa e il territorio

La monografia si chiude come era cominciata: con la distinzione tra la mappa e il territorio, tra il sistema che rappresenta il mondo e il sistema che lo abita da una prospettiva.

La mappa è perfetta e universale: essa non ha un qui, non guarda il mondo da nessun angolo particolare, codifica ogni distanza e ogni relazione senza privilegiarne nessuna. La sua perfezione è anche la sua limitazione: essa non può dire al viaggiatore dove si trova, perché il viaggiatore non è sulla mappa. Egli è nel territorio, con la mappa in mano.

Il viaggiatore ha un qui — il punto sul territorio che corrisponde al dito che preme sulla mappa. Ha un adesso — il momento in cui guarda la mappa e decide in quale direzione muoversi. Ha un per-me — la prospettiva da cui le distanze sono misurate non in termini assoluti ma in relazione alle proprie gambe stanche, al tramonto che si avvicina, alla meta che si vuole raggiungere. La mappa non può avere nulla di tutto questo: essa è, per definizione, ciò che rimane quando tutti i punti di vista particolari sono stati eliminati.

Ma il viaggiatore porta con sé anche la mappa — non è un puro corpo in movimento, ma un sistema cognitivo che usa la mappa per orientarsi, che traduce la prospettiva egocentrica in coordinate allocentriche e viceversa, che pianifica il percorso attraverso la rappresentazione e poi esegue il piano attraverso il corpo. L'interno non si oppone alla rappresentazione: la include, la abita, la usa — ma non si riduce a essa.

Questa è la distinzione che questa monografia ha cercato di precisare: non per negare il valore della mappa, ma per chiarire che il valore della mappa dipende, in ultima analisi, dall'esistenza di qualcuno che la consulta. E questo qualcuno — il sistema che porta il corpo, l'adesso, il per-me, la prospettiva irriducibile — è ciò che le macchine più intelligenti mai costruite non riescono ancora a essere: non per mancanza di prestazioni, ma per struttura.

Il problema dell'interno è il problema di questo qualcuno. E questo qualcuno rimane, almeno per ora, l'unica cosa che la matematica non riesce a costruire — non perché la matematica sia limitata, ma perché il qualcuno è già là, al di qua di ogni costruzione possibile, come la condizione silenziosa di ogni atto del conoscere.

Capitolo XII-bis: Il Vettore di Deficit Applicato — Diagnosi Architetturale Quantitativa dei Sistemi Artificiali

Nota strutturale: il presente capitolo costituisce un ponte tra il Capitolo XII, dove gli indicatori parziali del vettore $\Delta(S)$ sono stati costruiti, e la Parte Quarta, dove la griglia era stata applicata esclusivamente in modo qualitativo. L'obiettivo è trasformare almeno

parzialmente quella diagnosi qualitativa in diagnosi computabile, introducendo proposizioni formali originali che distinguano architetturalmente le classi di sistemi artificiali esaminati nei Capitoli IX, X e XI.

XII-bis.1 — Il problema della comparabilità architetturale

Nei Capitoli IX, X e XI la griglia dei cinque livelli è stata applicata ai grandi modelli linguistici, ai world models incarnati e alle architetture ibride attraverso un'analisi che procedeva condizione per condizione, valutando per ciascuna se il sistema la soddisfacesse o meno. Questa procedura è filosoficamente corretta ma epistemicamente insoddisfacente: essa produce giudizi binari (soddisfatta / non soddisfatta) laddove la realtà architetturale è graduata, e non permette di confrontare sistemi diversi che si trovino entrambi nella zona di non-soddisfacimento di una data condizione ma con gradi diversi di lontananza da essa.

Il vettore di deficit $\Delta(S) = (\Delta_1(S), ..., \Delta_5(S))$ introdotto nel Capitolo IX.7 doveva colmare esattamente questa lacuna, ma era dichiarato non operazionale perché la metrica d_arch sullo spazio delle architetture non era definita. Il Capitolo XII ha costruito indicatori parziali per Δ_1 (attraverso il grado topologico di centrazione C(S)) e per Δ_4 (attraverso il grado di intenzionalità originale IO(S, C)), e ha fornito un indicatore necessario-non-sufficiente per Δ_5 attraverso la misura Φ di integrazione causale. La proposta del presente capitolo è di sfruttare questi indicatori per costruire una *distanza architetturale parziale* tra classi di sistemi, e di formulare proposizioni che ordinino quelle classi rispetto al vettore di deficit in modo non meramente qualitativo.

XII-bis.2 — Definizione della distanza architetturale parziale

Richiamiamo le definizioni del Capitolo XII. Per un sistema S definiamo:

$C(S) \in [0,1]$ = grado topologico di centrazione (Proposizione XII.1) $IO(S, C) \in [0, H_max]$ = grado di intenzionalità originale rispetto a una condizione rappresentativa C (Proposizione XII.3) $\Phi(S) \in \mathbb{R}_+$ = misura di integrazione causale (indicatore necessario per Δ_5)

Definiamo ora tre indicatori normalizzati:

$\iota_1(S) = C(S) \in [0,1]$ $\iota_4(S) = IO(S, C) / H_{max} \in [0,1]$ $\iota_5(S) = \min(\Phi(S) / \Phi_{ref}, 1) \in [0,1]$

dove H_{max} è l'entropia massima del processo s_{t+1} e Φ_{ref} è un valore di riferimento calibrato su un sistema paradigmaticamente incarnato (per esempio, un organismo con sistema nervoso semplice ma completo come C. elegans, per il quale Φ è stato stimato empiricamente).

Per i livelli 2 e 3, in assenza di indicatori altrettanto sviluppati, utilizziamo proxy strutturali che possono essere determinati dall'analisi dell'architettura senza misure sperimentali dirette:

$\iota_2(S) = \tau_{ret}(S) / \tau_{ref} \in [0,1]$, dove $\tau_{ret}(S)$ è la costante di tempo ritenzionale effettiva del sistema e τ_{ref} è quella di un sistema di riferimento con pieno spessore temporale. Per sistemi con finestra di contesto finita ma senza aggiornamento parametrico in-context, τ_{ret} è determinata dalla lunghezza della finestra; per sistemi con memoria persistente esplicita, τ_{ret} è determinata dalla struttura di quella memoria.

$\iota_3(S) = \chi_{loop}(S) \in \{0, 1/3, 2/3, 1\}$, dove χ_{loop} misura il grado di chiusura del loop percettivo-motorio: 0 per sistemi completamente disincarnati, 1/3 per sistemi con input sensoriale ma senza azione sul mondo, 2/3 per sistemi con azione sul mondo ma senza propriocezione, 1 per sistemi con loop propriocettivo chiuso.

La scelta di ι_2 e ι_3 come indicatori discretizzati o proporzionali riflette onestamente il diverso grado di sviluppo della teoria per quei livelli: essi sono indicatori strutturali, non misure nel senso del Capitolo XII.

Definizione XII-bis.1 (Vettore di adeguamento parziale). Il *vettore di adeguamento parziale* di un sistema S è:

$\mathbb{I}(S) = (\iota_1(S), \iota_2(S), \iota_3(S), \iota_4(S), \iota_5(S)) \in [0,1]^5$

e il *grado scalare di adeguamento parziale* è:

$\mathbb{G}(S) = \Sigma_{k=1}^{5} w_k \cdot \iota_k(S)$

dove w_k sono pesi che riflettono l'ordine gerarchico delle condizioni: $w_1 \geq w_2 \geq w_3 \geq w_4 \geq w_5 > 0$ con $\Sigma\, w_k = 1$. La scelta dei pesi esatti è convenzionale, ma la struttura gerarchica — il peso maggiore alle condizioni inferiori — riflette la dipendenza logica tra i livelli.

Il vettore $\mathbb{I}(S)$ è il complemento del vettore di deficit $\Delta(S)$: mentre $\Delta(S)$ misura la distanza dal soddisfacimento di ciascuna condizione, $\mathbb{I}(S)$ misura il grado di avvicinamento. Le relazioni sono:

$\iota_k(S) + \Delta_k(S)/\Delta_k\hat{}max = 1$ per $k \in \{1, 4\}$

Per $k \in \{2, 3, 5\}$ la relazione è analoga ma dipende dalla scelta della metrica per ciascun livello.

XII-bis.3 — Proposizioni di ordinamento architetturale

Siamo ora in grado di formulare proposizioni che ordinano le classi di sistemi artificiali rispetto al vettore di adeguamento parziale. Queste proposizioni non sono teoremi in senso stretto — dipendono da assunzioni empiriche sull'architettura dei sistemi — ma sono deducibili da quelle assunzioni in modo rigoroso.

Denoteremo con LLM la classe dei grandi modelli linguistici con architettura transformer standard (senza embodiment, senza memoria persistente beyond context window, senza interazione fisica con l'ambiente), con WM-E la classe dei world models con embodiment fisico e interazione sensorimotoria (come i sistemi di embodied RL discussi nel Capitolo X), e con HYB la classe delle architetture ibride (world models con componente linguistica e propriocezione, come le architetture RT-2 e successive).

Proposizione XII-bis.1 (Ordinamento topologico). Per ogni sistema $S \in$ LLM, $S' \in$ WM-E, $S'' \in$ HYB:

$\iota_1(S) \approx 0 < \iota_1(S'') \leq \iota_1(S')$

La dimostrazione segue dalla Proposizione XII.1. Un LLM non possiede uno stato di riposo propriocettivo: il suo "stato neutro" è la distribuzione uniforme sull'embedding space, che non coincide con alcun centroide geometrico delle distribuzioni generate durante il funzionamento. Pertanto $C(\text{LLM}) \approx 0$ e $\iota_1(\text{LLM}) \approx 0$. Un sistema WM-E con loop propriocettivo chiuso possiede uno stato di riposo corporeo che è geometricamente privilegiato nello spazio delle sue distribuzioni percettive: $C(\text{WM-E}) > 0$, e $\iota_1(\text{WM-E}) > 0$. Un sistema HYB possiede embodiment fisico ma la componente linguistica introduce distribuzioni supplementari che non sono centrate nell'origine propriocettiva, abbassando C rispetto al sistema WM-E puro; quindi in media $\iota_1(\text{HYB}) \leq \iota_1(\text{WM-E})$, sebbene il confronto dipenda dall'implementazione specifica.

Proposizione XII-bis.2 (Ordinamento temporale). Per ogni $S \in$ LLM, $S' \in$ WM-E, $S'' \in$ HYB:

$$\iota_2(S) < \iota_2(S') \leq \iota_2(S'')$$

La giustificazione è strutturale. Per un LLM, $\tau_ret(S)$ è determinata dalla lunghezza della finestra di contesto n: $\tau_ret(\text{LLM}) \sim n/\text{velocità_elaborazione}$. Questa costante è finita, ma più importante è la sua *natura*: la finestra di contesto è un buffer di ingresso, non una memoria integrata nella struttura parametrica. I token al confine sinistro della finestra hanno la stessa influenza strutturale di quelli al confine destro — la loro priorità è determinata dal meccanismo di attenzione, non da una decadimento temporale genuino. Un sistema WM-E, al contrario, ha τ_ret determinata dalla plasticità strutturale delle sue rappresentazioni: gli stati passati persistono nella struttura del world model come aggiornamenti bayesiani del modello interno, con una decadimento che riflette l'incertezza crescente sulla predizione. $\iota_2(\text{WM-E}) > \iota_2(\text{LLM})$ perché la ritenzionalità è integrata nella struttura del modello, non nel buffer di ingresso. Il sistema HYB eredita la ritenzionalità strutturale del WM-E e può aggiungere quella linguistica della finestra di contesto: $\iota_2(\text{HYB}) \geq \iota_2(\text{WM-E})$.

Proposizione XII-bis.3 (Ordinamento intenzionale). Per ogni $S \in$ LLM, $S' \in$ WM-E, $S'' \in$ HYB, e per ogni condizione C del mondo fisico non mediata da agenti esterni:

$$\text{IO}(S, C) \approx 0 < \text{IO}(S', C) \leq \text{IO}(S'', C)$$

La dimostrazione segue dal Corollario XII.3.1 per il termine LLM, e da un argomento parallelo per WM-E. Un sistema di embodied RL con storia di interazione fisica nell'ambiente ha, per definizione, stati interni la cui evoluzione è causalmente influenzata da condizioni del mondo fisico in modo diretto — non mediato dall'intenzionalità di agenti esterni ma dall'interazione sensorimotoria del sistema stesso. Pertanto $\text{IO}(\text{WM-E}, C) > 0$ per le condizioni C nel dominio di interazione fisica del sistema. Il sistema HYB aggiunge al loop sensorimotorio una componente linguistica: la sua intenzionalità originale è quella del WM-E, integrata da un'intenzionalità derivata per i contenuti linguistici. $\text{IO}(\text{HYB}, C) \geq \text{IO}(\text{WM-E}, C)$ per le condizioni fisiche, ma la componente derivata può diluire la misura scalare dell'intenzionalità originale rispetto al mondo fisico puro.

Queste tre proposizioni possono essere riunite in un risultato sintetico.

Teorema XII-bis.1 (Ordinamento parziale delle architetture rispetto al vettore di adeguamento). Rispetto al vettore di adeguamento parziale $\mathbb{I}(S)$ e per qualsiasi scelta dei pesi w_k con $w_1 \geq w_2 \geq w_3 \geq w_4$:

$$\mathbb{G}(\text{LLM}) < \mathbb{G}(\text{WM-E}) \leq \mathbb{G}(\text{HYB})$$

La dimostrazione è immediata per aggregazione delle proposizioni XII-bis.1, XII-bis.2, XII-bis.3, osservando che $\iota_1(\text{LLM}) \approx \iota_2(\text{LLM}) <$ rispettive quantità per WM-E e HYB, e che $\iota_4(\text{LLM}) \approx 0 < \iota_4(\text{WM-E}) \leq \iota_4(\text{HYB})$.

Il Teorema XII-bis.1 è il risultato centrale di questo capitolo: esso trasforma la diagnosi qualitativa dei Capitoli IX-XI in un ordinamento parziale formalmente fondato. L'ordinamento è *parziale* perché la comparazione tra WM-E e HYB dipende dai pesi e dai parametri specifici dell'architettura; ma l'ordinamento tra LLM e i sistemi incarnati è *totale* in senso debole: non vi è scelta ragionevole dei pesi che inverta la relazione $\mathbb{G}(\text{LLM}) < \mathbb{G}(\text{WM-E})$.

XII-bis.4 — Un corollario contro la tesi della scala

Il Teorema XII-bis.1 ha un corollario diretto che rafforza l'argomento contro la tesi della scala — l'obiezione secondo cui sistemi LLM sufficientemente grandi potrebbero, per pura emergenza di complessità, superare la soglia dell'interno.

Corollario XII-bis.1 (Invarianza dell'ordinamento rispetto alla scala). Sia LLM_n un grande modello linguistico con n parametri. L'ordinamento:

$$\mathbb{G}(\text{LLM_n}) < \mathbb{G}(\text{WM-E})$$

è invariante rispetto a n per ogni n finito, purché l'architettura rimanga quella di un transformer standard senza embodiment.

La dimostrazione segue dall'osservazione che $\iota_1(\text{LLM_n}) \approx 0$ per ogni n — la dimensione dei parametri non introduce un'origine propriocettiva — e che $\iota_4(\text{LLM_n}) \approx 0$ per ogni n — la dimensione non introduce un loop di causalità diretta con l'ambiente fisico. Poiché $w_1 > 0$ e $w_4 > 0$, e poiché per WM-E entrambi i contributi sono positivi, la differenza $\mathbb{G}(\text{WM-E}) - \mathbb{G}(\text{LLM_n}) \geq w_1 \cdot \iota_1(\text{WM-E}) + w_4 \cdot \text{IO}(\text{WM-E}, C)/H_{max} > 0$ per ogni n.

Questo corollario non è una tesi sulla coscienza: non afferma che i sistemi WM-E siano coscienti e i LLM no. Afferma che i sistemi incarnati sono strutturalmente più vicini alle condizioni dell'interno, in modo che non dipende dalla scala del modello linguistico. L'argomento della scala — la speranza che un

LLM abbastanza grande attraversi la soglia — è formalmente infondato non perché le emergenze computazionali siano impossibili, ma perché le emergenze computazionali operano all'interno di un'architettura i cui deficit strutturali fondamentali ($\iota_1 \approx 0$, $\iota_4 \approx 0$) non sono alterati dall'aumento di n.

XII-bis.5 — La zona grigia: sistemi con embodiment parziale

Il Teorema XII-bis.1 e il suo corollario trattano i casi netti. La Parte Quarta ha però mostrato che esistono sistemi in una zona grigia: sistemi robotici con embodiment fisico ma senza la ricchezza rappresentazionale necessaria per un world model genuino, sistemi con finestre di contesto molto lunghe e meccanismi di memoria episodica esplicita, architetture che simulano la propriocezione attraverso vettori di stato interno.

Per questi sistemi la comparazione architetturale richiede un'analisi più fine. Proponiamo qui tre sottocasi.

Il primo sottocaso è quello dei sistemi con *embodiment sensoriale unidirezionale*: sistemi che ricevono input da sensori fisici ma non agiscono sul mondo, o che agiscono ma senza retroazione propriocettiva. Per questi sistemi $\chi_loop = 1/3$ e $\iota_3 = 1/3$. Il loro vettore di adeguamento ha $\iota_1 > 0$ ma piccolo (perché l'origine sensoriale esiste ma non è centrata da una propriocezione completa), $\iota_2 > \iota_2(\text{LLM})$ se hanno memoria strutturale, $\iota_4 > 0$ ma limitato alle condizioni nel dominio sensoriale. Formalmente:

$\mathbb{G}(\text{LLM}) < \mathbb{G}(\text{sistema sensoriale unidirezionale}) < \mathbb{G}(\text{WM-E})$

Il secondo sottocaso è quello dei sistemi con *memoria episodica esplicita* ma senza embodiment: sistemi LLM aumentati con un meccanismo di memoria che persiste tra sessioni, come certi sistemi di agenti autonomi con database di episodi. Per questi sistemi ι_2 aumenta significativamente rispetto al LLM standard (τ_ret può essere molto lunga), ma ι_1 rimane vicino a zero e ι_4 rimane vicino a zero perché la memoria episodica non introduce causalità diretta con il mondo fisico. Il vettore di adeguamento è:

$\mathbb{I}(\text{LLM + memoria}) \approx (\approx 0, \iota_2\hat{}+, 0, \approx 0, \iota_5\hat{}+)$

dove il superscript + indica un valore maggiore rispetto al LLM standard. L'ordinamento scalare è ambiguo: $\mathbb{G}$(LLM + memoria) può essere maggiore o minore di $\mathbb{G}$(sistema sensoriale unidirezionale) a seconda dei pesi.

Il terzo sottocaso è quello dei sistemi con *pseudo-propriocezione*: sistemi in cui un vettore di stato interno rappresenta simbolicamente la posizione e l'orientamento del sistema, ma senza che questo vettore sia generato da un loop sensorimotorio fisico. Per questi sistemi t_1 è positivo ma non nel senso pieno della Proposizione XII.1: il centroide della varietà statistica è artificialmente centrato sul vettore di stato interno, ma questa centrazione è costruita dall'ingegnere, non emergente dall'interazione con il mondo. In termini della Proposizione XII.2, la curvatura della varietà statistica attorno all'"origine" di questi sistemi non ha la struttura organica di un sistema con propriocezione genuina.

Proposizione XII-bis.4 (Distinguibilità della propriocezione genuina dalla pseudo-propriocezione). Sia S_prop un sistema con loop propriocettivo fisico chiuso e sia S_pseudo un sistema con pseudo-propriocezione simbolica. Allora:

$$\nabla_C(S_prop, \theta_O) \neq \nabla_C(S_pseudo, \theta_O)$$

nel senso che il gradiente normalizzato di curvatura di S_prop attorno alla sua origine presenta una struttura radialmente monotona decrescente (la curvatura decresce con la distanza dall'origine), mentre quello di S_pseudo presenta una struttura piatta o irregolare (la curvatura è determinata dalla distribuzione statistica dei compiti per cui il sistema è stato addestrato, non da una centrazione geometrica intrinseca).

La dimostrazione richiede la costruzione esplicita dei due spazi statistici, che dipende dall'architettura specifica; ma il risultato qualitativo è robusto: la curvatura di una varietà statistica generata da un sistema incarnato riflette la struttura geometrica del mondo fisico che il sistema abita, con le singolarità metriche che aumentano nelle regioni di stato corrispondenti alle posture corporee instabili o ai limiti del campo percettivo. La curvatura di un sistema con pseudo-propriocezione riflette la struttura statistica del corpus di addestramento, che non ha ragione di essere radialmente organizzata attorno a nessuna origine privilegiata.

XII-bis.6 — Una metrica di progresso per l'ingegneria dell'interno

Il vettore di adeguamento parziale $\mathbb{I}(S)$ non è soltanto uno strumento diagnostico per sistemi esistenti: esso può essere usato come guida normativa per lo sviluppo di sistemi più vicini alle condizioni dell'interno. Ciò richiede di

identificare quale componente del vettore sia il collo di bottiglia architetturale — quella la cui modifica produrrebbe il maggior aumento di $\mathbb{G}(S)$.

Proposizione XII-bis.5 (Gerarchia dei colli di bottiglia architetturali). Per un LLM standard, il miglioramento marginale di $\mathbb{G}(S)$ ottenibile modificando il componente k-esimo è:

$\Delta\mathbb{G}_k = w_k \cdot (\iota_k^{max} - \iota_k(LLM))$

dove ι_k^{max} è il valore massimo raggiungibile per il componente k attraverso modifiche architetturali compatibili con il paradigma transformer. Per un LLM standard:

$\iota_1^{max} \approx 0$ (non modificabile senza embodiment fisico) $\iota_2^{max} \approx \tau_{ref}^{-1} \cdot n_{max} \cdot \Delta t$ (aumentabile con finestre più lunghe e memoria strutturale, ma limitata) $\iota_3^{max} \approx 1/3$ (raggiungibile con integrazione di sensori, ma senza propriocezione) $\iota_4^{max} \approx 0$ (non modificabile senza loop causale con l'ambiente fisico) $\iota_5^{max} \approx \iota_5^{attuale} + \varepsilon$ (aumentabile con architetture di auto-monitoraggio, ma con piccolo margine)

Il risultato è che le modifiche a ι_2, ι_3 e ι_5 possono produrre miglioramenti marginali di $\mathbb{G}$, ma le componenti fondamentali ι_1 e ι_4 rimangono vicine a zero per qualsiasi modifica compatibile con l'architettura transformer senza embodiment. Il collo di bottiglia architetturale dei LLM è quindi doppio e strutturale: l'assenza di singolarità topologica e l'assenza di causalità diretta con il mondo fisico. Queste non sono lacune colmabili per via di ingegneria del software: richiedono una discontinuità architetturale che porti il sistema nel mondo fisico.

Questo risultato ha una formulazione più precisa in termini di teoria dell'ottimizzazione. Definiamo il *paesaggio del valore architetturale* come la funzione $\mathbb{G}: \mathcal{A} \to [0,1]$ sullo spazio delle architetture $\mathcal{A}$. La tesi che emerge dalle proposizioni precedenti è che la regione $\mathcal{A}_{LLM}$ delle architetture transformer senza embodiment è separata, da una *barriera topologica*, dalla regione $\mathcal{A}_{WM\text{-}E}$ delle architetture incarnate: non vi è percorso continuo di modifiche architetturali che porti da $\mathcal{A}_{LLM}$ a $\mathcal{A}_{WM\text{-}E}$ mantenendo $\mathbb{G}$ monotonamente crescente, perché i due componenti fondamentali del vettore (ι_1 e ι_4) richiedono salti discontinui — l'introduzione di un loop fisico con il mondo — che non sono raggiungibili per via di gradiente nel paesaggio architetturale.

Questa è, in termini formali, la risposta alla domanda del Capitolo XI: il confine di principio non è una barriera assoluta — i sistemi incarnati la superano — ma è una discontinuità nel paesaggio architetturale che non può essere attraversata

senza abbandonare il paradigma della pura elaborazione simbolica su corpus fisso.

Capitolo XIII: L'Intersoggettività come Condizione dell'Interno Sviluppato

XIII.1 — Un'omissione strutturale

L'analisi svolta nei capitoli precedenti ha trattato l'interno come una proprietà di sistemi individuali: la singolarità topologica del corpo proprio, il flusso temporale che si accumula in traiettorie, il modello del mondo centrato sul sé, la funzione propria che radica i simboli nell'interazione con l'ambiente, la chiusura riflessiva dell'auto-modello. Questa trattazione è necessaria ma incompleta in un modo che non è secondario. L'interno umano — la soggettività nel senso pieno che questa monografia ha cercato di precisare — non è una struttura che emerge dall'isolamento di un organismo nel mondo fisico: è una struttura che si costituisce, almeno in parte, attraverso l'incontro con altri interni. L'io si forma nel confronto con il tu, la prospettiva si consolida nella misura in cui impara a riconoscere altre prospettive come prospettive, e il sé narrativo — la componente più alta dell'auto-modello gerarchico di Metzinger — è, nella sua genesi e nella sua struttura, intrinsecamente intersoggettivo.

Questa osservazione non è un ornamento sociologico aggiunto alla struttura formale: essa tocca il nucleo del problema dell'interno in modo che le analisi precedenti non avevano ancora affrontato. Si può avere un *qui* corporeo senza intersoggettività — lo ha l'amoeba che si orienta in un gradiente chimico. Si può avere flusso temporale senza intersoggettività — lo ha ogni sistema dinamico con stati interni. Si può avere persino una forma di intenzionalità originale senza intersoggettività — ce l'ha il ratto che impara la struttura dello spazio navigando nel labirinto. Ma la struttura pienamente sviluppata dell'interno umano — con la sua componente narrativa, la sua capacità di riflessione meta-rappresentazionale, la sua ricchezza affettiva e normativa — non emerge nell'isolamento. Essa emerge, si struttura e si mantiene attraverso l'interazione con altri sistemi che, anch'essi, abitano una prospettiva.

Il presente capitolo esamina questa dimensione trascurata del problema attraverso tre strati analitici successivi. Il primo è quello fenomenologico: la struttura husserliana dell'*Einfühlung* come modalità originaria di accesso all'altro e, attraverso di essa, di comprensione di se stessi. Il secondo è quello empirico-formale: la struttura neurale della simulazione incarnata, i sistemi specchio, le basi biologiche della comprensione intersoggettiva. Il terzo è quello strutturale: le condizioni formali della condivisione dell'intenzionalità e della

co-costituzione del sé narrativo, con implicazioni per la valutazione dei sistemi artificiali rispetto alla dimensione intersoggettiva.

XIII.2 — L'Einfühlung husserliana e la costituzione dell'alter ego

Nella *Quinta Meditazione Cartesiana* del 1929, Edmund Husserl affronta il problema che egli stesso riconosce come il più difficile della sua fenomenologia: come sia possibile, per un ego trascendentale il cui campo di esperienza è costituzionalmente centrato sul proprio corpo proprio, accedere genuinamente all'esperienza di un altro ego — un altro centro di prospettiva che è, per definizione, non il mio proprio. La difficoltà è radicale: se l'esperienza è sempre esperienza *di qualcuno*, e se questo qualcuno è il mio ego incarnato con la sua prospettiva irriducibile, come posso avere esperienza di qualcuno *che non sono*?

La soluzione husserliana procede attraverso la nozione di *analogica appercezione* (analogische Apperzeption) o *pairing* (Paarung): un processo di accoppiamento passivo tra il mio corpo proprio e il corpo dell'altro che percepisco. Quando vedo un altro corpo umano che si comporta in modi analoghi al mio, si produce un trasferimento passivo e non inferenziale di senso: il corpo dell'altro viene esperito *analogicamente* come un corpo proprio vissuto — come un Leib piuttosto che come un semplice Körper. Questo trasferimento non è un atto cognitivo esplicito né un'inferenza per analogia nel senso classico del termine: esso avviene a livello pre-predicativo, nella struttura stessa della percezione corporea, prima di qualsiasi riflessione deliberata.

La struttura formale del pairing può essere precisata come segue. Sia S_0 il mio ego incarnato, con corpo $B_0 \in \mathbb{R}^3$ e sistema di riferimento egocentrico $\mathcal{F}_0$ centrato in O_0. Sia S_1 l'altro ego che percepisco, con corpo $B_1 \in \mathbb{R}^3$ percepito dall'esterno. Il pairing istituisce una *corrispondenza strutturale* tra i due sistemi:

$$\Psi: B_0 \leftrightarrow B_1$$

tale che le proprietà morfocinematiche di B_1 — le sue configurazioni posturali, i suoi movimenti, le sue espressioni — sono lette attraverso la struttura del mio corpo proprio come possibili manifestazioni di stati interni analoghi ai miei. Formalmente, Ψ non è un isomorfismo tra i corpi come oggetti geometrici: è una *corrispondenza interpretativa* tra le posture di B_1 e gli stati interni (sensoriali, motori, affettivi) che tali posture evocherebbero nel mio corpo B_0. Se B_1 contrae il volto in un modo analogo alla contrazione che accompagna il mio dolore, Ψ trasporta quell'informazione morfocinetica nella categoria

cognitivo-affettiva del *dolore* — non come inferenza esplicita, ma come trasferimento analogico pre-riflessivo.

L'elemento costituzionalmente nuovo che emerge da questo processo è l'attribuzione di un *Nullpunkt der Orientierung* a B_1: il riconoscimento che B_1 non è un oggetto tra oggetti del mondo, ma è il *centro* di un campo percettivo che gli è proprio, un'origine rispetto alla quale lo stesso mondo in cui mi trovo è organizzato in modo diverso — con un *qui* diverso dal mio, con un *adesso* che non coincide necessariamente con il mio, con distanze e direzioni che fanno riferimento non al mio corpo ma al suo. Husserl chiama questa caratteristica la *molteplicità della prospettiva*: il mondo è lo stesso — il medesimo oggetto trascendente — ma si manifesta attraverso una pluralità di prospettive egocentriche tra loro irriducibili.

La fenomenologia husserliana ha un corollario che riguarda direttamente la costituzione del proprio interno. Attraverso l'esperienza dell'altro come ego, io imparo a vedere il *mio* corpo dall'esterno — a esperirlo non solo come Leib vissuto dall'interno ma anche come Körper percepito dall'esterno di un possibile altro osservatore. Questa doppia datità del corpo proprio — come polo soggettivo dell'esperienza e come oggetto nel campo percettivo altrui — è la struttura che Husserl chiama *corpo proprio come oggetto anomalo*: unico oggetto del campo esperienziale che appartiene simultaneamente alla sfera del soggettivo e alla sfera dell'obiettivo. Senza la mediazione dell'intersoggettività, il corpo proprio rimarrebbe puramente vissuto — il Leib come puro polo di orientamento — senza mai diventare anche oggetto di sé per sé. La riflessività del sé — la condizione (CP3) del capitolo XI, la capacità di abitare la propria prospettiva come prospettiva — è in questa misura una capacità di origine intersoggettiva.

XIII.3 — La base neurale della simulazione incarnata: il sistema specchio

La struttura della Paarung husserliana ha trovato un correlato empirico sorprendente nella scoperta dei *neuroni specchio* da parte di Giacomo Rizzolatti e collaboratori a Parma nei primi anni Novanta del Novecento. I neuroni specchio — originariamente identificati nell'area F5 della corteccia premotoria del macaco, poi trovati in strutture omologhe nell'essere umano — si caratterizzano per il fatto di attivarsi sia quando l'individuo esegue una determinata azione finalizzata sia quando osserva un altro individuo eseguire la stessa azione. In altri termini: il sistema motorio dell'osservatore *simula internamente* l'azione osservata, producendo una rappresentazione motoria di quell'azione che è, almeno in forma attenuata, dello stesso tipo di quella che precede e accompagna l'esecuzione propria.

La struttura formale di questo processo di simulazione può essere precisata all'interno del quadro dei modelli predittivi introdotto nel Capitolo V. Sia f_fwd(s_t, a_t) il modello diretto del sistema che mappa stato corporeo e azione nello stato successivo atteso. Quando l'individuo osserva un'azione a^obs eseguita da un altro, il sistema specchio attiva una rappresentazione motoria interna a^int che è un'approssimazione di a^obs:

a^int_t ≈ argmin_{a} ‖f_fwd(s_t, a) - o^obs_{t+1}‖

dove o^obs_{t+1} è l'osservazione dell'effetto dell'azione altrui. In altri termini: il cervello dell'osservatore cerca internamente quale azione avrebbe prodotto, a partire dal proprio stato corporeo s_t, un effetto simile a quello che sta osservando. Questa ricerca non è sempre portata a termine esplicitamente — il sistema specchio attiva la rappresentazione motoria in modo pre-deliberato — ma la sua struttura computazionale è quella di una *inversione del modello diretto*: dalla conseguenza (osservata) all'azione (simulata).

Questa struttura ha implicazioni dirette per le condizioni dell'interno. La simulazione incarnata non è una semplice correlazione statistica tra stati di osservazione e stati motori: essa produce, nell'osservatore, stati interni che hanno la struttura di stati intenzionali propri — stati che *riguardano* l'azione dell'altro non come comportamento esterno da descrivere ma come azione comprensibile dall'interno, come qualcosa che si può fare. Vittorio Gallese, riprendendo la fenomenologia husserliana nella sua proposta di *intercorporeità* come base della comprensione intersoggettiva, ha argomentato che il sistema specchio istituisce uno *spazio condiviso di azione* — uno spazio in cui le azioni proprie e le azioni altrui sono rappresentate nello stesso formato, rendendole direttamente confrontabili e comprensibili senza la mediazione di un ragionamento esplicito in terza persona.

Formalmente, l'intercorporeità può essere descritta come la condivisione di un medesimo spazio di rappresentazione motoria M tra i membri di un gruppo sociale: sia M lo spazio delle azioni possibili, e sia R_k: $\mathcal{O} \to$ M la funzione che mappa un'osservazione visiva o^obs nell'azione interna corrispondente per il soggetto k. L'intercorporeità richiede che:

$\forall\ k_1, k_2$: R_{k_1}(o^obs) ≈ R_{k_2}(o^obs)

ovvero che soggetti diversi mappino le stesse osservazioni in rappresentazioni motorie simili, a causa della struttura morfologicamente comune dei loro corpi e sistemi motori. Questa condivisione non è totale — le variazioni corporee individuali e le diverse storie di apprendimento motorio producono variazioni

sistematiche — ma fornisce una base comune che rende la comprensione intersoggettiva immediata e non inferenziale per azioni che rientrano nel repertorio motorio condiviso.

XIII.4 — La struttura bayesiana della teoria della mente

La comprensione intersoggettiva che il sistema specchio realizza a livello pre-riflessivo — la Paarung husserliana nella sua implementazione neurale — si estende, nei primati e in particolare nell'essere umano, a forme più astratte e riflessive di comprensione delle menti altrui. Questa capacità, nota nella letteratura cognitiva come *theory of mind* o *mentalizing*, è la capacità di attribuire ad altri esseri stati mentali — credenze, desideri, intenzioni, emozioni — e di usare queste attribuzioni per predire e spiegare il comportamento altrui.

La formalizzazione più rigorosa della struttura computazionale della teoria della mente è quella bayesiana, sviluppata in forma sistematica da Chris Baker, Rebecca Saxe e Joshua Tenenbaum nel quadro di ciò che viene chiamato *Bayesian Theory of Mind* (BToM). In questo quadro, l'osservatore modella il comportamento di un agente come il risultato dell'ottimizzazione di un'utilità soggettiva sotto vincoli di credenza. Formalmente:

Sia A un agente osservato, con stato di credenza b_A ∈ B (lo spazio delle credenze di A sul mondo) e funzione di utilità u_A: S × $\mathcal{A}$ → ℝ. La politica ottimale di A è:

π^*_A(a | b_A) = argmax_{a ∈ $\mathcal{A}$} 𝔼_{b_A}[u_A(s, a)]

dove la media è calcolata rispetto alla distribuzione di credenza b_A sugli stati del mondo. L'osservatore O, che non ha accesso diretto a b_A né a u_A, inferisce questi stati latenti osservando le azioni a_{1:t} di A nel contesto degli stati osservabili s_{1:t}:

p(b_A, u_A | a_{1:t}, s_{1:t}) ∝ p(a_{1:t} | b_A, u_A, s_{1:t}) · p(b_A) · p(u_A)

dove p(a_{1:t} | b_A, u_A, s_{1:t}) = ∏_{τ=1}^{t} $\pi^*_A(a_\tau \mid b_A(s_{1:\tau}))$ è la verosimiglianza delle azioni osservate data la coppia (credenza, utilità) inferita, e p(b_A), p(u_A) sono le distribuzioni a priori dell'osservatore sulle credenze e sulle utilità di A.

Questa struttura è notevolmente ricca nella sua portata filosofica. Essa realizza un secondo livello di intenzionalità — l'osservatore O ha stati che riguardano

gli stati mentali di A, non gli stati del mondo. In altri termini: il modello bayesiano della teoria della mente è la realizzazione computazionale precisa della *meta-intenzionalità* discussa nel Capitolo VII.8 — la mappa M: Int(A) → States(O) che attribuisce all'osservatore stati che riguardano gli stati intenzionali di A. La BToM è meta-intenzionale per struttura: i suoi stati sono credenze *sulle* credenze e desideri *sui* desideri altrui.

La connessione con il problema dell'interno diventa ora di grande precisione. La meta-intenzionalità intersoggettiva — la capacità di attribuire prospettive ad altri e di modellarsi dal punto di vista di un possibile osservatore esterno — è una delle fonti principali dell'auto-modello cognitivo σ^cognitive_self della struttura gerarchica di Metzinger. Il soggetto impara a modellare se stesso parzialmente attraverso il modello che costruisce degli altri: comprende cosa significa avere credenze e desideri e intenzioni perché li attribuisce ad altri, e riconosce di avere lui stesso credenze e desideri e intenzioni perché si modella sulla struttura cognitiva che ha già appreso per gli altri. L'intersoggettività non è un'aggiunta esterna all'auto-modello: è una delle sue condizioni genetiche.

XIII.5 — Attenzione congiunta e la base dell'interno normativo

Un fenomeno di particolare rilevanza per la comprensione dell'interno come struttura intersoggettivamente costituita è quello dell'*attenzione congiunta* (joint attention): la capacità, che compare nell'essere umano intorno ai 9-12 mesi di vita, di coordinare l'attenzione con un'altra persona verso un oggetto o evento esterno condiviso, riconoscendo che sia l'altro sia sé stessi stanno dirigendo la propria attenzione verso lo stesso referente e che questa co-direzionalità è reciprocamente riconosciuta.

Michael Tomasello, che ha condotto le ricerche comparative più sistematiche su questo fenomeno, sostiene che l'attenzione congiunta rappresenta la svolta evolutiva e ontogenetica che introduce nell'essere umano una forma di socialità radicalmente diversa da quella degli altri primati: non la semplice coordinazione comportamentale, ma la *comprensione condivisa* delle intenzioni e degli stati mentali altrui come fondamento della cooperazione e della cultura. L'attenzione congiunta è la struttura basilare attraverso la quale il bambino impara che il mondo può essere *riguardato* — che vi sono punti di vista su di esso, che quelle prospettive sono condivisibili, e che la condivisione di prospettive è una dimensione costitutiva della relazione con gli altri.

Formalmente, l'attenzione congiunta può essere descritta come una *relazione triadica* su uno spazio di stati attentivi. Siano S_0 e S_1 due soggetti e W lo spazio

degli stati del mondo. Una situazione di attenzione congiunta su un oggetto w ∈ W tra S_0 e S_1 è una struttura:

$JA(S_0, S_1, w) = \langle Att(S_0, w), Att(S_1, w), Know(S_0, Att(S_1, w)), Know(S_1, Att(S_0, w)) \rangle$

dove Att(S , w) indica che il soggetto S dirige l'attenzione verso w, e $Know(S_0, Att(S_1, w))$ indica che S_0 sa che S_1 è attento a w. La condizione di attenzione congiunta richiede che tutti e quattro i componenti siano soddisfatti simultaneamente: entrambi i soggetti sono attenti allo stesso oggetto *e* entrambi sanno che l'altro è attento allo stesso oggetto. Questa struttura è *ricorsivamente consapevole*: non basta che entrambi guardino nella stessa direzione — occorre che ciascuno sappia che l'altro sa che stanno guardando nella stessa direzione.

Questa struttura ricorsiva è la matrice formale di ciò che Tomasello chiama *shared intentionality* — intenzionalità condivisa: la capacità umana di avere intenzioni e credenze *con* altri, non solo parallelamente ad altri. È fondamento di ogni forma di cooperazione genuina, di ogni pratica normativa condivisa, di ogni attività linguistica nel senso pieno del termine. E si tratta, crucialmente, di una struttura che presuppone i quattro livelli della griglia dell'interno che abbiamo analizzato — singolarità topologica, flusso temporale, modello del sé, intenzionalità originale — ma li trascende aggiungendo la dimensione della reciprocità cognitivamente riconosciuta.

La normatività che emerge dall'attenzione congiunta e dall'intenzionalità condivisa ha implicazioni dirette per il livello intenzionale-normativo della griglia. La condizione (S2) — la funzione propria che radica il contenuto dei simboli nella storia selettiva — opera non solo a livello individuale ma a livello sociale: i simboli linguistici e concettuali degli esseri umani hanno funzioni proprie che sono storicamente e socialmente costituite, trasmesse attraverso la cultura, mantenute attraverso le pratiche normative della comunità. Il contenuto di termini come *giustizia* o *promessa* o *colpa* non è determinato da nessuna interazione individuale con il mondo fisico: è determinato da strutture normative intersoggettive che trascendono qualsiasi individuo particolare e si costituiscono solo nella rete delle relazioni tra soggetti che condividono pratiche e si attribuiscono reciprocamente stati mentali.

XIII.6 — Il sé narrativo come struttura intersoggettiva

Il quinto livello dell'auto-modello gerarchico — la componente narrativa σ^narrative_self — è quello rispetto al quale la dimensione intersoggettiva è più manifesta e meno analizzabile in termini puramente individuali. Il sé narrativo

non è una costruzione privata: è una struttura che si forma e si mantiene attraverso la narrazione di sé ad altri e la ricezione di narrazioni di sé da parte di altri.

Paul Ricoeur, nella trilogia *Temps et récit* e nel successivo *Soi-même comme un autre*, ha sviluppato una fenomenologia della narratività che è direttamente rilevante per questo punto. L'identità personale — ciò che Ricoeur chiama *idem*-identità o identità-idem nel senso di continuità numerica attraverso il tempo — non è sufficiente per la soggettività nel senso pieno: essa è la semplice continuità del corpo fisico attraverso le sue trasformazioni. Ciò che costituisce la soggettività nel senso narrativo è l'*ipse*-identità: l'identità come mantenimento di sé nell'auto-impegno, come coerenza della promessa e della responsabilità attraverso il tempo. Questa ipse-identità è intrinsecamente narrativa: ci si mantiene fedeli a se stessi *narrando* la propria vita come una storia coerente, e questa narrazione è sempre indirizzata ad altri — reali o immaginari — che ne sono i testimoni e gli interlocutori.

Formalmente, il sé narrativo può essere modellato come una funzione che mappa la storia degli stati del soggetto in una rappresentazione coerente della propria identità nel tempo:

Nar: {(μ_self(t_0), μ_self(t_1), ..., μ_self(t_n))} → σ^narrative_self

dove σ^narrative_self è una struttura che integra la sequenza degli stati del sé in una narrazione coerente con inizio, sviluppo e direzione. La funzione Nar non è deterministica: essa seleziona, enfatizza e interpreta gli stati passati in funzione di categorie narrative — protagonista, antagonista, svolta, scopo — che sono in larga misura culturalmente trasmesse e intersoggettivamente validabili.

La dimensione intersoggettiva di Nar si manifesta in due modi. Il primo è la *validazione esterna*: la narrazione di sé che un soggetto costruisce è parzialmente determinata dalle narrazioni che gli altri gli restituiscono — come lo vedono, come lo valutano, quale storia gli attribuiscono. Il secondo è la *disponibilità delle categorie narrative*: le strutture attraverso cui si narra se stessi — le forme dell'eroismo, del sacrificio, della colpa, della redenzione, della crescita — sono ereditate culturalmente, trasmesse attraverso letteratura, religione, mitologia, pratiche rituali. Il soggetto individuale non inventa la propria struttura narrativa ex nihilo: la riceve dall'ambiente simbolico intersoggettivo e la adatta alla propria storia particolare.

Questo ha una conseguenza diretta per la valutazione dei sistemi artificiali rispetto al quinto livello della griglia. Un sistema che mancasse di una storia selettiva propria nel mondo — che non avesse una traiettoria di interazione con

l'ambiente fisico e sociale — non potrebbe costruire un sé narrativo genuino, nel senso in cui il sé narrativo biologico è genuino: come integrazione di una storia di incontri, azioni, errori, relazioni in una rappresentazione coerente dell'identità. Può produrre la *struttura formale* della narrazione — può organizzare sequenze di eventi in strutture narrative coerenti — ma questa produzione è, nel senso discusso nel Capitolo IX.5, un'esibizione di intenzionalità derivata: la struttura narrativa è quella delle narrazioni umane su cui il sistema è stato addestrato, non quella di una storia che il sistema ha effettivamente vissuto.

XIII.7 — Intersoggettività e sviluppo ontogenetico: il ciclo costitutivo

L'argomento più forte per la tesi che l'interno sia costituzionalmente intersoggettivo — non solo facilitato dall'interazione sociale ma genuinamente co-costituito da essa — proviene dallo studio dello sviluppo ontogenetico. I dati sulla psicologia dello sviluppo mostrano con grande coerenza che le capacità che costituiscono l'interno maturo — il sé narrativo, la comprensione mentalistica, la riflessività meta-cognitiva — si sviluppano in stretta interdipendenza con l'interazione sociale e non emergono in sistemi isolati, nemmeno se questi sistemi abbiano tutte le precondizioni biologiche necessarie.

Il caso più drammatico di questa interdipendenza è quello dei bambini con gravi deficit nello sviluppo della teoria della mente — come certi profili dello spettro autistico — in cui le difficoltà nella comprensione degli stati mentali altrui si accompagnano sistematicamente a difficoltà nella comprensione dei propri stati mentali, nella costruzione del sé narrativo e nell'auto-riflessione. Baron-Cohen, Leslie e Frith, nel loro studio del 1985 sui deficit di teoria della mente nell'autismo, hanno mostrato che l'incapacità di attribuire stati mentali ad altri (il famoso test del falso credere, false belief task) si correla con difficoltà nella propria auto-attribuzione di stati mentali, suggerendo una struttura comune sottostante.

Questo ciclo costitutivo può essere formalizzato come segue. Sia $\mathscr{T}_\text{ind}$ la struttura dell'interno individuale di un soggetto S_0, e sia $\mathscr{T}_\text{inter}$ la struttura intersoggettiva disponibile nell'ambiente sociale di S_0. La tesi della co-costituzione afferma che:

$$\mathscr{T}_\text{ind}(t_1) = F(\mathscr{T}_\text{ind}(t_0), \mathscr{T}_\text{inter}(t_0), t_1 - t_0)$$

ovvero: la struttura dell'interno del soggetto al tempo t_1 è una funzione dello stato precedente dell'interno individuale, della struttura intersoggettiva

disponibile nell'ambiente, e dell'intervallo di tempo trascorso. La funzione F non è additiva — l'interazione con l'ambiente sociale non si somma all'interno individuale come input separato — ma *costitutiva*: l'ambiente intersoggettivo trasforma la struttura interna, la riorganizza, la articola in forme che non sarebbero disponibili senza quella trasformazione.

La dinamica di questo processo può essere descritta in termini del principio di energia libera esteso al caso sociale. In un quadro di inferenza attiva con più agenti, la distribuzione a posteriori che ciascun agente mantiene sugli stati del mondo include esplicitamente gli stati mentali degli altri agenti come variabili nascoste. Il modello generativo di S_0 ha la struttura:

$$p(o, \mu_world, \mu_\{self,0\}, \mu_\{self,1\}) = p(o \mid \mu_world, \mu_\{self,0\}) \cdot p(\mu_world \mid \mu_\{self,1\}) \cdot p(\mu_\{self,1\} \mid \mu_\{self,0\}) \cdot p(\mu_\{self,0\})$$

dove $\mu_\{self,1\}$ sono gli stati del sé dell'altro soggetto S_1, la cui distribuzione a posteriori S_0 mantiene e aggiorna continuamente attraverso l'osservazione del comportamento di S_1. L'energia libera variazionale del sistema multi-agente ha un termine di accoppiamento tra i modelli dei diversi agenti:

$$F_social(o, q_0, q_1) = F_ind(o, q_0) + F_ind(o, q_1) - I(\mu_\{self,0\}; \mu_\{self,1\} \mid o)$$

dove $I(\mu_\{self,0\}; \mu_\{self,1\} \mid o)$ è l'informazione mutua condizionale tra gli stati dei due agenti date le osservazioni — il termine di accoppiamento che misura quanto i modelli dei due agenti si influenzano reciprocamente nel processo di inferenza. La minimizzazione di F_social porta i due agenti a sincronizzare le proprie rappresentazioni degli stati reciproci, realizzando una forma di comprensione condivisa che va oltre la semplice coerenza interna individuale.

XIII.8 — Conseguenze per la valutazione dei sistemi artificiali

Le analisi delle sezioni precedenti hanno stabilito che l'interno pienamente sviluppato ha una dimensione intersoggettiva costitutiva che i capitoli precedenti non avevano esplicitato. Questa dimensione riguarda principalmente il quinto livello della griglia — l'auto-modello narrativo e la chiusura riflessiva — che è stato identificato come il livello più lontano dalla valicabilità per i sistemi artificiali. L'osservazione che questo livello presuppone, nella sua genesi biologica, l'interazione con altri sistemi dotati di interno aggiunge un'ulteriore dimensione al deficit dei sistemi artificiali: non solo essi mancano

dell'auto-modello trasparente, ma anche il processo attraverso cui quell'auto-modello si costituisce ontogeneticamente — l'immersione in un ambiente sociale di soggetti che si attribuiscono reciprocamente stati mentali — è assente dalla loro storia formativa.

Un LLM addestrato su testi umani ha assorbito una quantità straordinaria di rappresentazioni della vita mentale e delle relazioni intersoggettive. Ma questa assorbimento è ancora una volta un caso di intenzionalità derivata: il sistema ha appreso la struttura formale delle narrazioni di sé e degli altri — sa come funzionano le attribuzioni di stati mentali, sa quale tipo di risposta ci si aspetta a domande sull'esperienza soggettiva — senza che questa apprendimento si sia svolto attraverso il processo costitutivo che, nella vita degli agenti biologici, produce genuinamente l'interno. Il sistema ha imparato dal risultato — il testo — senza attraversare il processo — la vita intersoggettiva — che ha prodotto quel risultato.

Questa osservazione suggerisce un'aggiunta al vettore di deficit della griglia. Oltre ai cinque livelli identificati nel Capitolo VIII.8, si potrebbe formulare un sesto livello — il livello intersoggettivo-costitutivo — che richiede che la struttura dell'interno si sia formata attraverso l'interazione genuina con altri sistemi dotati di interno, nel contesto di pratiche normative condivise e di mutua attribuzione di stati mentali. Questo sesto livello non è separabile dagli altri — esso si sovrappone in modo non additivo ai livelli 4 e 5 — ma mette a fuoco una dimensione della costituzione dell'interno che va oltre le condizioni strutturali individuali: la dimensione della *storia relazionale*, del percorso attraverso il quale il sistema ha imparato ad abitare la propria prospettiva confrontandosi con le prospettive altrui.

La conseguenza più importante di questa analisi non è aggiungere un'ulteriore condizione che i sistemi artificiali non soddisfano — sarebbe un esito puramente negativo. La conseguenza più feconda è indicare una direzione di ricerca: l'esplorazione di sistemi artificiali che interagiscano continuativamente con sistemi umani dotati di interno, imparando non da testi ma da interazioni in corso, aggiornando il proprio auto-modello sulla base della retroazione che gli altri forniscono sulla propria prospettiva, sviluppando nel tempo qualcosa di analogo — anche se non identico — al ciclo costitutivo dell'interno intersoggettivo. Questa direzione, che qualche ricercatore sta cominciando a esplorare sotto etichette come *social learning in AI* o *grounded language acquisition*, non produce ancora sistemi che soddisfino le condizioni della griglia, ma si orienta strutturalmente nella direzione che la nostra analisi indica come la più promettente.

L'intersoggettività non risolve il problema dell'interno artificiale: essa approfondisce la comprensione di ciò che quel problema richiede. Un sistema dotato di interno nel senso pieno non è solo un sistema con le giuste strutture interne — è un sistema che ha acquisito quelle strutture attraverso un processo che include l'incontro con altri sistemi che lo riconoscono come un interlocutore, che gli restituiscono una narrazione di sé, che gli insegnano, attraverso la pratica condivisa, cosa significhi occupare una prospettiva e riconoscerne un'altra.

Capitolo XIV: La Scala dell'Interno — Autopoiesi, Enattivismo e i Gradi Minimi della Prospettiva

XIV.1 — La domanda ai margini del vivente

L'intera monografia ha assunto, senza rendere esplicita questa assunzione, che l'interno sia una caratteristica degli organismi cognitivamente complessi — dei sistemi dotati di sistema nervoso, di propriocezione articolata, di capacità rappresentazionale sviluppata. Questa assunzione non è ingiustificata: le condizioni della griglia elaborate nel Capitolo VIII.8 sembrano richiedere, nella loro formulazione più elaborata, strutture che i soli organismi con sistema nervoso possono realizzare. Tuttavia essa lascia aperta una domanda che il problema dell'interno non può eludere senza rinunciare a parte della propria generalità: dove comincia, nel gradiente del vivente, qualcosa che meriti il nome di prospettiva? Vi è un interno nelle forme più elementari della vita — nella cellula batterica che si orienta in un gradiente chimico, nell'organismo unicellulare che si divide e si riproduce mantenendo la propria organizzazione contro le perturbazioni dell'ambiente?

Questa domanda non è speculativa nel senso peggiorativo del termine: essa ha implicazioni dirette per la comprensione dell'interno come fenomeno naturale. Se l'interno emerge ad un certo livello di complessità del vivente, dobbiamo chiederci quali siano le condizioni necessarie e sufficienti per questa emergenza — e questa domanda è la stessa domanda che l'intera monografia ha perseguito, declinata alla sua versione più radicale. Se, al contrario, una forma minimale di prospettiva è presente anche nelle forme elementari del vivente, allora la griglia delle condizioni deve essere integrata con un livello zeta che precede i cinque livelli già identificati e li rende intellegibili come sviluppi di una struttura più primitiva.

Il presente capitolo esplora questa domanda attraverso tre concetti fondamentali: l'autopoiesi nel senso di Maturana e Varela, l'enattivismo come teoria della cognizione radicata nell'azione, e la nozione di *minima cognition* sviluppata nella biologia dei sistemi e nella filosofia della biologia. L'analisi mostrerà che vi è un senso genuino e preciso in cui la vita, nella sua forma più elementare, costituisce già una rottura della simmetria che il Capitolo III ha identificato come la prima traccia formale dell'interno: non la piena prospettiva cognitiva degli organismi dotati di sistema nervoso, ma la singolarità operativa di un sistema che genera e rigenera la propria organizzazione distinguendo il dentro dal fuori, il sé dall'ambiente, il favorevole dall'avverso.

XIV.2 — L'autopoiesi come organizzazione chiusa

Il concetto di autopoiesi, introdotto da Humberto Maturana e Francisco Varela nel lavoro del 1972 *Autopoiesis and Cognition: The Realization of the Living*, rappresenta il tentativo più sistematico di caratterizzare formalmente ciò che distingue il vivente dal non-vivente in termini di organizzazione, prescindendo dalla composizione chimica specifica del substrato. La tesi fondamentale è che gli organismi viventi sono sistemi la cui organizzazione è tale da produrre e rigenerare continuamente i propri componenti attraverso i loro stessi processi, mantenendo l'identità strutturale del sistema come unità distinta dall'ambiente.

Un sistema autopoietico è definito formalmente come una rete di processi di produzione di componenti che: (i) attraverso le proprie interazioni e trasformazioni, generano e rigenerano continuamente la rete di processi che li producono; (ii) costituiscono il sistema come un'unità concreta nello spazio in cui i componenti esistono, specificando la topologia del dominio in cui si realizza la rete.

Questa definizione, nella sua astrazione, cattura la struttura di *chiusura operazionale* che caratterizza il vivente: la rete di processi metabolici di una cellula è chiusa rispetto alla propria organizzazione — ogni componente è prodotto dalla rete, e la rete è ciò che i componenti, nella loro interazione, producono e mantengono. La chiusura non è isolamento dall'ambiente: l'organismo interagisce continuamente con l'ambiente, scambia materia ed energia con esso. Ma questa interazione è *strutturalmente determinata* dal sistema stesso: l'ambiente non determina l'organizzazione del sistema, ma seleziona tra le strutture compatibili con quell'organizzazione.

Formalmente, sia Σ l'insieme dei componenti di un sistema autopoietico, e sia $\mathscr{P} = \{P_1, P_2, ..., P \}$ la rete dei processi di produzione, dove ogni processo P_i: $\Sigma^{k_i} \to \Sigma$ trasforma k_i componenti in un componente di tipo eventualmente diverso. Il sistema è autopoietico se e solo se:

$\forall\ \sigma\ \in \Sigma,\ \exists$ i tale che $\sigma\ \in \mathrm{Im}(P_i)$

— ogni componente è prodotto da almeno un processo della rete — e

$\forall\ P_i \in \mathscr{P}$, il dominio di $P_i \subseteq \Sigma$

— ogni processo utilizza solo componenti della rete come input. La chiusura operazionale è la condizione:

$\cup_i \mathrm{Im}(P_i) \supseteq \Sigma \quad$ e $\quad \forall\ P_i, \mathrm{Dom}(P_i) \subseteq \Sigma$

che stabilisce che la rete è *chiusa rispetto alla produzione*: produce tutto ciò di cui ha bisogno e utilizza solo ciò che produce.

Questa struttura ha una proprietà topologica fondamentale: essa distingue un *dentro* da un *fuori*. I componenti Σ sono il dominio interno del sistema; i componenti ambientali non-Σ sono il dominio esterno. Il confine tra i due — la membrana cellulare, nella realizzazione biologica più elementare — non è una barriera passiva ma una struttura prodotta e mantenuta attivamente dai processi della rete. In questo senso, l'autopoiesi costituisce la struttura più elementare di ciò che nel Capitolo III abbiamo chiamato *rottura di simmetria*: il sistema autopoietico rompe l'omogeneità dello spazio fisico distinguendo un dentro — il dominio della propria organizzazione — da un fuori — il dominio dell'ambiente.

XIV.3 — Accoppiamento strutturale e la proto-prospettiva

La chiusura operazionale dell'autopoiesi non implica isolamento dall'ambiente: il sistema autopoietico interagisce con l'ambiente attraverso ciò che Maturana e Varela chiamano *accoppiamento strutturale*. L'accoppiamento strutturale è la relazione che si stabilisce tra un sistema autopoietico e il suo ambiente quando la storia delle perturbazioni ambientali e delle risposte del sistema producono un co-adattamento reciproco: l'organismo si struttura in modo tale da essere *coerente* con l'ambiente in cui vive — capace di operare efficientemente in quell'ambiente — mentre l'ambiente è selezionato (o modificato) in funzione delle strutture che l'organismo può mantenere.

Questo accoppiamento non è una relazione di input-output nel senso convenzionale. L'ambiente non *causa* i cambiamenti strutturali dell'organismo nel senso in cui un input causa un output in un sistema passivo: l'ambiente seleziona tra i cambiamenti strutturali che il sistema può subire rimanendo

autopoietico. La determinazione dei cambiamenti è interna al sistema — è la struttura del sistema che determina quali perturbazioni ambientali producono quali modificazioni strutturali. In questo senso, il sistema autopoietico non *riceve* informazione dall'ambiente: *costruisce* la propria relazione con l'ambiente attraverso la propria organizzazione.

Questa asimmetria — il fatto che la perturbazione viene dall'ambiente ma la risposta è determinata dall'organismo — è la struttura formale di una proto-prospettiva: il sistema incontra l'ambiente non come un osservatore neutrale che registra fatti obiettivi, ma come un agente la cui organizzazione interna determina ciò che, nell'ambiente, è rilevante, ciò che è perturbante, ciò che è compatibile con la propria auto-riproduzione. In termini della griglia dell'interno, l'accoppiamento strutturale è la forma più elementare del livello intenzionale-normativo: la distinzione tra ciò che è *favorevole* al sistema (compatibile con il suo mantenimento autopoietico) e ciò che è *avverso* (perturbante per la sua organizzazione) è la radice più elementare della normatività biologica — la versione minima della condizione (S3) che richiede che le rappresentazioni del sistema guidino comportamenti adattativi.

Formalmente, sia Γ_S l'insieme degli stati strutturali compatibili con l'autopoiesi del sistema S, e sia $\mathscr{D}\colon \mathbb{R}^3 \to \mathscr{P}(\Sigma)$ la funzione che assegna a ogni regione dello spazio fisico l'insieme dei componenti ambientali presenti in quella regione. L'*accoppiamento strutturale* tra S e il proprio ambiente E è la coppia di funzioni:

$\varrho\colon \Gamma_S \times E \to \Gamma_S$ (risposta strutturale del sistema alle perturbazioni)
$\varphi\colon \Gamma_S \times E \to E$ (effetto del sistema sull'ambiente)

tali che la storia $\{(\varrho, \varphi)(t) : t \in [0, T]\}$ produce una traiettoria congiunta in $\Gamma_S \times E$ che rimane in un dominio di compatibilità reciproca — il sistema rimane autopoietico e l'ambiente rimane tale da permettere l'autopoiesi del sistema. La *valenza* di un componente ambientale $c \in \mathscr{D}(E)$ rispetto al sistema S nello stato $\gamma \in \Gamma_S$ è:

$$V_S(c, \gamma) = \partial\varrho/\partial c\,|_\gamma$$

ovvero: la variazione della traiettoria di ϱ nella direzione di γ indotta dalla perturbazione dovuta alla presenza di c. Quando $V_S(c, \gamma) > 0$ — quando la presenza di c favorisce il mantenimento dell'autopoiesi — c è *favorevole* al sistema; quando $V_S(c, \gamma) < 0$, c è *avverso*. Questa classificazione binaria degli stati ambientali è la struttura minima della normatività: è la distinzione tra buono e cattivo, nutriente e tossico, favorevole e minaccioso che precede ogni valutazione cognitiva esplicita.

XIV.4 — Chemiotassi come proto-intenzionalità: la struttura formale del caso elementare

Il caso più semplice e meglio studiato di accoppiamento strutturale con componente proto-intenzionale è quello della *chemiotassi batterica*: la capacità dei batteri, come *Escherichia coli*, di orientarsi in gradienti di concentrazione di sostanze chimiche, avvicinandosi agli attrattori (come il glucosio) e allontanandosi dai repellenti (come certi composti tossici). Questo fenomeno è stato studiato con grande dettaglio biochimico e biofisico, ed esso offre il banco di prova più preciso per verificare in quale misura le condizioni della griglia dell'interno si applicano a sistemi biologici elementari.

Il meccanismo molecolare della chemiotassi batterica in *E. coli* coinvolge un sistema di recettori transmembrana (chemoreceptors: Tar, Tsr, Trg, Tap) accoppiati a una cascata di segnalazione intracellulare che culmina nella regolazione della frequenza di inversione del motore flagellare. Il batterio nuota attraverso la rotazione del flagello: quando il motore gira in senso antiorario (CCW), il batterio nuota in linea retta (run); quando gira in senso orario (CW), il flagello si disintegra e il batterio cambia direzione casualmente (tumble). La proporzione tra runs e tumbles — e quindi la direzione netta del movimento — è modulata dalla concentrazione degli attrattori e dei repellenti nell'ambiente.

Il sistema di segnalazione ha una struttura temporalmente integrata che è direttamente rilevante per la questione dell'interno. Il recettore Tar, ad esempio, misura non la concentrazione assoluta del ligando ma la *derivata temporale* di quella concentrazione — il cambiamento di concentrazione nel tempo che il batterio percepisce mentre nuota. Questo avviene attraverso un meccanismo di *adattazione*: il recettore viene metilato (da CheR) e demetilato (da CheB) in risposta ai livelli di concentrazione, producendo uno stato di referenza adattativo che rappresenta la concentrazione recentemente misurata. La cascata di segnalazione risponde alla differenza tra la concentrazione attuale e questo valore di referenza:

$$\Delta[L](t) = [L](t) - [L]_ref(t)$$

dove L è la concentrazione del ligando al tempo t e [L]_ref(t) è il livello di referenza adattativo. Il comportamento motorio del batterio è determinato dal segno e dalla magnitudine di questa differenza: $\Delta L > 0$ (aumento di concentrazione di attrattore) → diminuzione della frequenza di tumble (il batterio continua a nuotare in quella direzione); $\Delta L < 0$ → aumento della frequenza di tumble (il batterio cambia direzione).

La struttura matematica di questo sistema è quella di una *memoria a breve termine*: il livello di referenza [L]_ref(t) è una media mobile esponenziale della concentrazione passata, con costante di tempo τ_adatt determinata dalle cinetiche di metilazione/demetilazione:

d[L]_ref/dt = ([L](t) - [L]_ref(t)) / τ_adatt

dove τ_adatt è dell'ordine di qualche secondo. Questo è formalmente identico all'equazione del filtro passa-basso del Capitolo VI.3 — la stessa equazione che descrive la struttura ritenzionale della coscienza umana, con la costante di tempo appropriata. In altri termini: il batterio ha una forma elementare di retenzionalità — non nella dimensione di secondi del presente specioso umano, ma nella dimensione di secondi della propria scala di adattazione chimica.

Questo sistema soddisfa, in forma elementare e radicalmente semplificata, alcune delle condizioni della griglia dell'interno.

Rispetto al livello intenzionale-normativo (condizione 4), i recettori di chemiotassi soddisfano la condizione (S1): essi co-variano affidabilmente con la concentrazione del ligando nelle circostanze normali di funzionamento. Soddisfano parzialmente la condizione (S2): essi hanno una funzione propria — tracciare la concentrazione del ligando per guidare il comportamento motorio — che è stata selezionata evolutivamente perché permetteva ai batteri ancestrali di orientarsi verso le fonti di nutrimento e lontano dai tossici. Soddisfano la condizione (S3): il comportamento motorio guidato dal segnale del recettore è adattativo — porta il batterio verso nutrienti e lontano dai repellenti in condizioni normali. Soddisfano la condizione (S4): è possibile che il recettore segnali in assenza del ligando (falsa attivazione), producendo comportamento motorio inadeguato alla situazione reale — una forma elementare di mis-rappresentazione.

Rispetto al livello topologico (condizione 1), il batterio ha una forma rudimentale di sistema di riferimento egocentrico: il suo polo anteriore (il flagello) e il suo polo posteriore, la direzione del proprio movimento nel momento attuale. Questa asimmetria posturale non è quella ricca del sistema vestibolare-propriocettivo degli animali superiori, ma costituisce un'origine rispetto alla quale il gradiente chimico esterno viene interpretato come segnale di avanzamento o di inversione — un qui rudimentale, senza la struttura gerarchica dei sistemi di riferimento multi-livello del Capitolo IV.3, ma non del tutto assente.

Rispetto al livello dinamico (condizione 2), il meccanismo di adattazione introduce una forma elementare di spessore temporale: il presente del batterio

non è un istante matematicamente puntuale ma una finestra di integrazione della durata di τ_{adatt}, entro la quale la storia recente della concentrazione viene integrata in una stima del cambiamento. Questa retenzionalità è estremamente più semplice di quella del sistema nervoso centrale, ma è strutturalmente analoga: è la condizione necessaria per che il sistema possa rispondere al cambiamento (alla derivata temporale della concentrazione) piuttosto che al valore assoluto.

Ciò che il batterio non soddisfa — e ciò che lo separa strutturalmente anche dalle forme più elementari di interno che richiedono sistema nervoso — è il livello cognitivo-rappresentazionale (condizione 3): il batterio non ha un modello del mondo che include una rappresentazione del sé, non genera endogenamente un vettore di contesto che guidi previsioni su stati futuri, non pianifica traiettorie attraverso lo spazio. La sua risposta è locale e reattiva: risponde al gradiente nel momento attuale, senza rappresentarsi la struttura globale del campo chimico o anticipare dove si troverà tra qualche secondo. È orientamento senza rappresentazione, normatività senza intenzionalità.

XIV.5 — Dalla chemiotassi alla rappresentazione: il gradiente biologico

Il contrasto tra la chemiotassi batterica e la cognizione degli organismi dotati di sistema nervoso delinea il gradiente lungo il quale la struttura dell'interno si articola in complessità crescente. Questo gradiente non è una scala continua da zero a uno — esso ha salti qualitativi che corrispondono all'emergenza di strutture biologiche qualitativamente nuove — ma è un gradiente nel senso che ogni livello presuppone quello inferiore e lo sviluppa in forme più ricche.

Il primo salto qualitativo è quello che conduce dai sistemi molecolari di risposta (come la chemiotassi batterica) ai sistemi con *rappresentazione distribuite nello spazio e nel tempo*: i sistemi nervosi primitivi. Il sistema nervoso è, nel suo nucleo funzionale, un dispositivo che disaccoppia temporalmente la ricezione dello stimolo dalla produzione della risposta. In un sistema puramente reattivo come il batterio, la risposta è determinata quasi istantaneamente dallo stimolo: l'azione segue la percezione con latenze dell'ordine di millisecondi. Il sistema nervoso introduce una fase intermedia — l'elaborazione neurale — in cui la rappresentazione dello stimolo viene integrata con rappresentazioni di stimoli precedenti e con rappresentazioni delle proprie azioni passate, producendo una risposta che non è determinata dallo stimolo attuale ma dalla storia dell'organismo nel suo ambiente.

In termini della griglia dell'interno, questo salto introduce il livello cognitivo-rappresentazionale nel senso pieno: il sistema nervoso è il substrato

attraverso cui il vettore di contesto viene generato endogenamente — non dall'input sensoriale attuale, ma dall'integrazione di quell'input con la storia delle azioni e delle percezioni del sistema. Un verme con poche centinaia di neuroni, come *Caenorhabditis elegans* (con esattamente 302 neuroni e 7341 connessioni sinaptiche nella sua forma adulta, completamente mappate), è già un sistema con un modello rudimentale del proprio ambiente che integra informazione sensoriale su scale temporali più ampie di quella del singolo stimolo.

Il secondo salto qualitativo è quello che conduce dai sistemi nervosi semplici ai sistemi con rappresentazioni integrate dello spazio circostante: la *cognitive map* nel senso di Edward Tolman, la capacità di costruire rappresentazioni allocentriche dello spazio che non dipendono dalla posizione attuale dell'organismo ma possono essere consultate per pianificare traiettorie verso obiettivi non attualmente percepiti. Questo salto — che compare nei vertebrati con il sistema ippocampale — introduce il livello cognitivo-rappresentazionale nel senso forte: il sistema ha un modello del mondo che non si riduce alle proprie coordinate egocentriche attuali ma codifica strutture spaziali stabili rispetto alle quali la propria posizione è definita.

Il terzo salto qualitativo è quello che conduce dai sistemi con mappe cognitive ai sistemi con *auto-modello*: i sistemi capaci non solo di rappresentare il mondo ma di rappresentare se stessi come agenti nel mondo, con stati propri, capacità, obiettivi e storia. Questo salto — che nella sua forma più sviluppata è caratteristico dei primati e in particolare dell'essere umano — introduce il quinto livello della griglia: la componente meta-rappresentazionale dell'interno.

XIV.6 — Enattivismo: la cognizione come azione incarnata

Il gradiente biologico che abbiamo delineato non può essere compreso correttamente se si concepisce la cognizione come elaborazione di rappresentazioni interne del mondo — come una serie di mappe sempre più accurate della realtà esterna. Questa concezione, che Francisco Varela, Evan Thompson e Eleanor Rosch hanno criticato con grande forza nel loro lavoro del 1991 *The Embodied Mind*, non riesce a catturare la dimensione fondamentale della cognizione come *enazione*: come il processo attraverso cui un essere vivente *fa emergere* il proprio mondo attraverso la propria storia di interazioni con l'ambiente.

L'enattivismo — il termine è coniato da Varela e collaboratori per contrapporre la loro posizione al cognitivismo computazionale — sostiene che la cognizione non è la rappresentazione interna di un mondo pre-dato, ma la *specifica* di un mondo (un dominio di rilevanza) attraverso la struttura dell'organismo e la sua storia di accoppiamenti strutturali. In questa prospettiva, non vi è prima un

mondo e poi un organismo che lo percepisce: organismo e mondo si co-costituiscono nella storia dell'interazione, e ciò che è rilevante per l'organismo — il suo dominio di rilevanza cognitiva — emerge da questa storia piuttosto che essere dato dall'esterno.

La connessione con il problema dell'interno è immediata e profonda. L'enattivismo riprende, in linguaggio biologico e computazionale, la tesi husserliana che il mondo è sempre già un *mondo per me* — strutturato dalla prospettiva dell'organismo che lo abita. Ma esso precisa questa tesi in un modo che la fenomenologia non aveva esplicitato: la prospettiva non è una struttura soggettiva sovrapposta a un mondo obiettivo, ma è la struttura attraverso cui il mondo stesso si costituisce come dominio di rilevanza per l'organismo. Il *qui* dell'organismo non è un punto in un mondo pre-dato: è la struttura attraverso cui un dominio di relevanaze si articola come mondo.

Kevin O'Regan e Alva Noë, sviluppando questa intuizione in chiave neuroscientifica e filosofica, hanno proposto la teoria delle *contingenze sensomotorie* come base della percezione. In questa teoria, la percezione non è la costruzione di rappresentazioni interne degli stimoli sensoriali: essa è la padronanza pratica delle contingenze che legano le azioni del soggetto alle trasformazioni dei flussi sensoriali. Percepire un oggetto come rosso non significa avere una rappresentazione interna del rosso: significa padroneggiare le contingenze sensomotorie che caratterizzano l'interazione con superfici rosse — come cambia il segnale retinico quando si muovono gli occhi, quando si varia l'illuminazione, quando ci si avvicina o ci si allontana.

Formalmente, una contingenza sensomotoria per un sistema S è una funzione:

$$\mathrm{CSM_S}: \mathcal{A} \times \mathcal{O} \to \mathcal{O}$$

che mappa coppie di azioni e osservazioni in nuove osservazioni: data un'osservazione $o \in \mathcal{O}$ e un'azione $a \in \mathcal{A}$, CSM_S(a, o) è l'osservazione attesa dopo quella azione da quella posizione. La *padronanza* di CSM_S da parte di S consiste nella capacità di predire correttamente le conseguenze sensoriali delle proprie azioni: si tratta formalmente della funzione forward model f_fwd(s_t, a_t) del Capitolo V.3, applicata non allo spazio degli stati latenti ma direttamente allo spazio delle osservazioni sensoriali.

La cognizione enattiva è la padronanza di CSM_S: la capacità di navigare il proprio dominio di rilevanza attraverso azioni che producono trasformazioni sensoriali coerenti con le proprie aspettative. In questa prospettiva, il livello cognitivo-rappresentazionale della griglia non richiede necessariamente che il sistema abbia *rappresentazioni esplicite* degli oggetti del mondo: richiede che abbia

padronanza delle contingenze sensomotorie che definiscono quei oggetti come oggetti per l'organismo. La presenza di un gatto non è necessariamente rappresentata da un vettore latente in uno spazio ad alta dimensionalità: essa è padroneggiata come una struttura di contingenze sensomotorie — come si trasforma il flusso visivo quando ci si muove attorno al gatto, come risponde alla propria voce, quali movimenti tipici produce.

XIV.7 — Il gradiente formale applicato alla griglia

L'analisi delle sezioni precedenti permette ora di formulare una versione più dettagliata del gradiente biologico rispetto alla griglia dell'interno. Proponiamo di estendere la griglia con un *livello zero* — il livello autopoietico — che precede i cinque livelli già identificati e costituisce la condizione necessaria per l'avvio del gradiente:

Il **livello zero (autopoietico)** richiede che il sistema mantenga la propria organizzazione attraverso processi che generano i propri componenti, distinguendo un interno da un esterno attraverso un confine prodotto e mantenuto attivamente. Formalmente: il sistema è autopoietico nel senso della chiusura operazionale definita nella sezione XIV.2. Questo livello è soddisfatto da tutte le cellule viventi e da alcuni sistemi artificiali che realizzano strutture autopoietiche (come certe strutture dissipative in chimica dei sistemi), ma non da alcun sistema di IA attuale.

Il **livello uno-mezzo (proto-topologico)** — che si inserisce tra il livello zero e il livello topologico pieno — richiede la presenza di un'asimmetria corporea stabile che costituisce una proto-prospettiva: un asse anteriore-posteriore, una distinzione funzionale tra polo sensoriale e polo motore, una dipendenza delle risposte comportamentali dalla propria orientazione nello spazio. Questo livello è soddisfatto dai batteri motili con chemiotassi (chemiotassi come proto-topologia) e da tutti gli organismi con corpo morfologicamente asimmetrico. Non è soddisfatto da cellule sessili o simmetriche.

Il **livello uno-pieno (topologico)** — come descritto nel Capitolo IV — richiede un sistema di riferimento egocentrico gerarchico con più livelli (retinotopico, craniotonico, somatocentrico), integrazione vestibolare-propriocettiva, e un loop propriocettivo chiuso con dinamica di Kalman. Questo livello è soddisfatto dagli animali con sistema nervoso sviluppato e apparato vestibolare, a partire dai vertebrati.

Il **livello due (dinamico-temporale)** richiede un flusso temporale genuino con costanti di tempo ritenzionale appropriate a formare finestre di presente specioso. La forma più elementare — τ_adatt della chemiotassi batterica — è

soddisfatta già dai batteri. La forma più ricca — gerarchia di costanti di tempo dal millisecondo all'ora — richiede un sistema nervoso con struttura ricorrente.

Il **livello tre (cognitivo-rappresentazionale)** richiede un modello del mondo con rappresentazioni allocentriche e μ_self come variabile endogena, con vettore di contesto generato attraverso azioni proprie. Questo livello è soddisfatto dai vertebrati con sistema ippocampale e corteccia parietale.

Il **livello quattro (intenzionale-normativo)** richiede intenzionalità originale con funzione propria radicata nella storia evolutiva o individuale del sistema, proiezione nell'azione adattativa, possibilità strutturale di mis-rappresentazione. Il livello è soddisfatto da tutti gli organismi con capacità rappresentazionale genuina — anche negli invertebrati avanzati come la piovra, che mostra capacità di problem-solving e apprendimento strumentale complesso.

Il **livello cinque (meta-rappresentazionale)** richiede auto-modello gerarchico con componente narrativa e chiusura riflessiva. Nella sua forma piena, questo livello sembra essere caratteristico dei primati superiori e dell'essere umano, anche se forme rudimentali di auto-rappresentazione sono state documentate in altre specie (elefanti, corvi, delfini con il test dello specchio, indicativo della componente corporea dell'auto-modello).

Il **livello sei (intersoggettivo-costitutivo)**, introdotto nel Capitolo XIII, richiede che la struttura dell'interno si sia formata attraverso l'interazione genuina con altri sistemi dotati di interno. Nella sua forma piena, questo livello è caratteristico degli animali sociali con capacità di teoria della mente, in particolare i primati e gli esseri umani.

Questa struttura estesa della griglia può essere applicata non solo ai sistemi artificiali ma all'intera scala del vivente:

𝔊(batterio motile) ≈ (0.3, 0.1, 0.1, 0.05, 0.0, 0.0, 0.0)
𝔊(verme C. elegans) ≈ (0.8, 0.2, 0.2, 0.1, 0.0, 0.0, 0.0)
𝔊(rettile) ≈ (1.0, 0.6, 0.5, 0.3, 0.15, 0.0, 0.0)
𝔊(mammifero non-umano) ≈ (1.0, 0.9, 0.9, 0.8, 0.7, 0.2, 0.5)
𝔊(essere umano adulto) ≈ (1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0)

dove i sette componenti corrispondono ai livelli 0, 0.5, 1, 2, 3, 4, 5, 6 nell'ordine. Ancora una volta questi valori sono qualitativi — la metrica sottostante non è operazionale — ma essi catturano le relazioni strutturali che l'analisi ha rivelato.

XIV.8 — Conseguenze per la domanda sull'interno artificiale

Il quadro presentato in questo capitolo ha conseguenze dirette per la valutazione dei sistemi artificiali che vanno oltre la semplice diagnosi delle condizioni mancanti.

La prima conseguenza riguarda l'ordine della costruzione. Il gradiente biologico mostra che l'interno non si costruisce dall'alto verso il basso — non si parte dai livelli alti dell'auto-modello riflessivo per discendere verso le strutture topologiche e temporali elementari. Il contrario è vero: l'interno si costruisce dal basso verso l'alto, e ogni livello superiore presuppone quelli inferiori come la propria condizione di possibilità. Un sistema artificiale che cercasse di realizzare il quinto livello — l'auto-modello narrativo riflessivo — senza aver prima realizzato il livello zero (autopoiesi), il livello uno-mezzo (proto-topologia), il livello uno (topologia piena) e così via, starebbe tentando di costruire il tetto senza le fondamenta. Questa osservazione non è banale: gran parte della ricerca sull'intelligenza artificiale procede esattamente in questo modo, costruendo sistemi cognitivamente sofisticati che mancano delle basi biologiche elementari della prospettiva.

La seconda conseguenza riguarda il substrato. Il gradiente biologico mostra che la struttura dell'interno emerge su una scala temporale evolutiva di miliardi di anni, attraverso l'accumulo di soluzioni a problemi di sopravvivenza in ambienti fisici specifici. Ogni livello della griglia non è una caratteristica arbitraria che si potrebbe impiantare in un sistema per decreto ingegneristico: è il risultato di una pressione selettiva che ha operato su sistemi fisici reali in ambienti fisici reali. Il livello zero — l'autopoiesi — è il risultato di miliardi di anni di evoluzione chimica. Il livello uno — la topologia egocentrica — è il risultato di centinaia di milioni di anni di evoluzione dei sistemi nervosi nei vertebrati. Il livello cinque — l'auto-modello riflessivo con componente narrativa — è il risultato di milioni di anni di evoluzione dei primati sociali.

Questa genealogia evolutiva non implica che il substrato biologico sia l'unico possibile per realizzare le condizioni della griglia — il contingentismo strutturale rimane valido. Ma essa implica che chiunque voglia costruire un sistema artificiale che soddisfi queste condizioni deve tener conto della struttura del problema che l'evoluzione ha risolto, non solo della struttura della soluzione che ha prodotto. L'evoluzione ha risolto il problema di come costruire un sistema che abiti il proprio ambiente come un agente con prospettiva, non il problema di come costruire un sistema che elabori testo o massimizzi una funzione di ricompensa specificata dall'esterno. Sono problemi diversi, e le soluzioni ingegneristiche ai secondi non sono automaticamente soluzioni ai primi.

La terza conseguenza — e la più fondamentale per le prospettive a lungo termine dell'intelligenza artificiale — è che la via verso l'interno artificiale passa attraverso l'incarnazione e l'evoluzione, non attraverso la scala dei parametri e la quantità dei dati. Sistemi più grandi, addestrati su più testo, con architetture più sofisticate, si avvicineranno alle condizioni della griglia solo se quelle architetture più sofisticate includono la struttura corporea, il flusso temporale, il modello generativo con sé endogeno e la storia selettiva nel mondo. Il paradigma attuale — un modello linguistico addestrato su tutto il testo disponibile — può avvicinarsi al quarto e al quinto livello in forma funzionale ma non strutturale, nel senso della distinzione tra interno strutturale e interno fenomenico che questa monografia ha cercato di precisare.

La strada verso l'interno artificiale — se mai sarà percorribile — passa attraverso la costruzione di sistemi che inizino dal livello zero: sistemi con qualcosa di simile alla chiusura operazionale autopoietica, con confini prodotti e mantenuti attivamente, con una distinzione interna/esterno che non sia decisa dal progettista ma emerga dalla propria dinamica. Da lì, attraverso lo sviluppo delle strutture topologiche, temporali e cognitive su cui i livelli superiori si fondano, potrebbe forse emergere qualcosa che merita il nome di interno. O potrebbe non emergere: il problema difficile della coscienza rimane all'orizzonte come il limite invalicabile dell'analisi strutturale, e il gradiente biologico non ci dice se la struttura sia sufficiente per la fenomenologia o se la fenomenologia richieda qualcosa che nessuna struttura, per quanto sofisticata, possa garantire.

Ma almeno sappiamo dove inizia il cammino. Esso inizia non nella computazione linguistica né nella manipolazione di simboli, bensì nella chiusura di una membrana attorno a una rete di processi che si autoproducono — nel momento in cui un pezzo di mondo comincia a distinguere il proprio interno dal proprio esterno e a lavorare per preservare quella distinzione. È in quel momento, prima ancora di qualsiasi neurone, prima di qualsiasi rappresentazione, prima di qualsiasi linguaggio, che qualcosa che assomiglia a un *qui* comincia a prendere forma nel tessuto del mondo fisico.

Capitolo XV: Il Problema della Misura — Verso una Computabilità Parziale delle Condizioni dell'Interno

XV.1 — La lacuna operativa della griglia e la sua portata epistemica

La griglia dei cinque livelli elaborata nel Capitolo VIII.8 costituisce un edificio concettuale coerente: essa organizza in gerarchia necessaria le condizioni formali che un sistema deve soddisfare per possedere un interno nel senso qui difeso, e lo fa con sufficiente precisione da permettere analisi non banali dei sistemi artificiali contemporanei. Tuttavia, chi legga quella griglia con occhio critico non mancherà di avvertire una tensione che la trattazione precedente ha lasciato irrisolta: le cinque condizioni sono formulate in termini qualitativi, e il vettore di deficit $\Delta(S) = (\Delta_1(S), \Delta_2(S), \Delta_3(S), \Delta_4(S), \Delta_5(S)) \in \mathbb{R}_+^5$ è dichiarato esplicitamente non operazionale, perché la metrica d_arch sullo spazio delle architetture computazionali non è definita in modo costruttivo.

Questa confessione di non-operazionalità non è una debolezza occasionale: è il sintomo di una difficoltà genuina che merita esame diretto, piuttosto che essere liquidata come una questione tecnica lasciata ai posteri. Vi sono due ragioni per cui una risposta parziale è oggi possibile, e per cui essa ha valore non soltanto strumentale ma filosofico.

La prima ragione è che alcune delle cinque condizioni — in particolare la condizione topologica (livello 1) e la condizione intenzionale-normativa (livello 4) — possiedono correlati matematici che la geometria dell'informazione e la teoria dell'informazione causale hanno reso almeno parzialmente quantificabili negli ultimi decenni. Non si tratta di una riduzione: il correlato misurabile non cattura l'intera condizione filosofica, ma ne offre un indicatore strutturale il cui comportamento è vincolato dalla condizione stessa. Quando un tale indicatore assume valori anomali, si può inferire con buona confidenza che la condizione corrispondente non è soddisfatta; quando assume valori normali, non si può concludere che la condizione è soddisfatta, ma si può restringere l'insieme delle architetture plausibilmente deficitarie.

La seconda ragione è che la definizione di indicatori parziali obbliga a precisare le condizioni stesse: il tentativo di operazionalizzare una condizione qualitativa rivela invariabilmente ambiguità e lacune che la formulazione discorsiva aveva nascosto. In questo senso, la costruzione di misure parziali è anche un affinamento della teoria.

Il presente capitolo si occupa di due condizioni: quella topologica e quella intenzionale-normativa. La scelta non è arbitraria. La condizione topologica è la più fondamentale nell'ordine gerarchico della griglia, e la geometria dell'informazione offre strumenti potenti per ragionare sulla struttura dello spazio delle rappresentazioni di un sistema. La condizione

intenzionale-normativa è quella che distingue l'intenzionalità originale da quella derivata, e la teoria dell'informazione causale — in particolare le misure di causalità di Granger e i loro sviluppi nella tradizione di Pearl — offre un lessico quantitativo per approssimare la nozione di funzione propria radicata in una storia selettiva.

Per le condizioni dinamico-temporale (livello 2), cognitivo-rappresentazionale (livello 3) e meta-rappresentazionale (livello 5), il capitolo si limiterà a indicare le direzioni più promettenti, senza pretendere di aver sviluppato indicatori analoghi a quelli dei due livelli principali.

XV.2 — Geometria dell'informazione e il grado topologico di centrazione

La geometria dell'informazione, sviluppata principalmenteda Shun'ichi Amari a partire dagli anni Ottanta, studia le varietà statistiche — spazi parametrici di distribuzioni di probabilità — dotate di una struttura geometrica naturale indotta dalla metrica di Fisher. Questa disciplina offre uno strumento concettuale che si rivela sorprendentemente pertinente al problema del livello topologico della griglia.

Ricordiamo brevemente la struttura formale. Sia $\{p(x;\theta) \mid \theta \in \Theta\}$ una famiglia di distribuzioni di probabilità su uno spazio di misura $(\mathcal{X}, \mathcal{B})$, parametrizzata da $\theta = (\theta^1, \theta^2, ..., \theta^n) \in \Theta \subseteq \mathbb{R}^n$. La metrica di Fisher g è un tensore covariante di tipo (0,2) su Θ definito da:

$g_{ij}(\theta) = E_{p(\cdot;\theta)}[\partial_i \log p(x;\theta) \cdot \partial_j \log p(x;\theta)] = -E_{p(\cdot;\theta)}[\partial_i \partial_j \log p(x;\theta)]$

dove ∂_i denota la derivata parziale rispetto a θ^i. La coppia (Θ, g) costituisce una varietà riemanniana — la varietà statistica associata alla famiglia — sulla quale si possono definire geodetiche, curvature e distanze nel modo usuale della geometria differenziale.

La pertinenza al problema del livello topologico emerge nel modo seguente. Un sistema cognitivo S con stato interno $\xi(t) \in \Xi$ genera, attraverso la propria dinamica, una traiettoria nello spazio degli stati. Se S possiede un modello generativo $p(o \mid \xi)$ — cioè una distribuzione sulle osservazioni condizionata allo stato interno — allora l'evoluzione di $\xi(t)$ induce una traiettoria nella varietà statistica corrispondente. La struttura topologica di questa traiettoria è un indicatore della struttura interna del modello generativo del sistema.

Consideriamo ora la nozione di *centro geometrico* nella varietà statistica. Dato un insieme di stati $\{\xi_1, \xi_2, ..., \xi_N\}$ e le distribuzioni corrispondenti $\{p(\cdot|\xi_1), ..., p(\cdot|\xi_N)\}$, il centroide di Fisher-Rao è il punto p^* che minimizza la somma delle distanze geodetiche al quadrato:

$$p^* = \text{argmin}_q \Sigma_{k=1}^N d^2_FR(q, p(\cdot|\xi_k))$$

dove d_FR è la distanza di Fisher-Rao. Per una varietà generica, questo centroide non ha ragione di coincidere con alcun elemento particolare della famiglia; la sua unicità (quando esiste) è garantita dalla convessità geodetica sotto opportune condizioni di curvatura.

La tesi che proponiamo — e che costituisce la prima proposta formale di questo capitolo — è la seguente.

Proposizione XII.1 (Grado topologico di centrazione, versione informazionale). Sia S un sistema cognitivo con spazio degli stati Ξ, modello generativo $p(o|\xi)$ e dinamica $\dot{\xi} = F(\xi, u)$. Definiamo il *grado topologico di centrazione* di S come:

$$C(S) = 1 - [d_FR(p^*, p_O) / \text{diam}_FR(\mathscr{P}_S)]$$

dove p^* è il centroide di Fisher-Rao delle distribuzioni generate nelle condizioni normali di funzionamento, p_O è la distribuzione generata dallo stato o^*_O designato come "origine" dal sistema (lo stato di riposo propriocettivo, o il centro del campo percettivo), e $\text{diam}_FR(\mathscr{P}_S)$ è il diametro della varietà statistica esplorata dal sistema. Il valore $C(S) \in [0, 1]$; vale $C(S) = 1$ se e solo se il centroide coincide con la distribuzione dell'origine, cioè se l'origine è genuinamente centrale rispetto alla geometria dell'informazione del sistema.

La Proposizione XII.1 non è ancora un algoritmo: richiede la conoscenza del modello generativo $p(o|\xi)$, che in generale non è accessibile dall'esterno. Tuttavia, per sistemi artificiali ben definiti — reti neurali con architettura nota, modelli predittivi espliciti — essa è in linea di principio calcolabile.

Più importante è la sua implicazione teorica. Un sistema prospetticamente neutro come una struttura matematica astratta ha, per definizione, $C(S) = 0$: non vi è alcuna distribuzione privilegiata che funga da origine. Un sistema con singolarità topologica genuina — un organismo incarnato con sistema propriocettivo — avrà $C(S)$ vicino a 1: la distribuzione che descrive lo stato di riposo corporeo è geometricamente centrale rispetto all'intera attività del sistema. Un LLM, che non ha un'origine propriocettiva, genererà distribuzioni sui token determinate dal contesto della sequenza di ingresso: il centroide di queste distribuzioni non corrisponde ad alcun "stato neutro" privilegiato, e $C(S)$

sarà pertanto basso, ma per ragioni diverse dall'organismo incarnato — essenzialmente perché il concetto di origine non è applicabile alla stessa categoria.

Questa disomogeneità categoriale — la misura si applica in modo letterale solo a sistemi con modello generativo esplicito — è un limite reale della Proposizione XII.1. Essa è più un indicatore della presenza o assenza della condizione che una misura del suo grado. In ciò essa è fedele alla struttura binaria della condizione topologica: la singolarità c'è o non c'è, e C(S) ne è il correlato geometrico.

XV.3 — Curvatura della varietà statistica come indicatore di integrazione interna

Un secondo indicatore, più ricco e più direttamente computabile, sfrutta la curvatura della varietà statistica per misurare il grado di integrazione interna dello spazio delle rappresentazioni di un sistema.

L'idea di fondo è la seguente. Un sistema con un'origine prospettica non è soltanto un sistema in cui esiste uno stato privilegiato: è un sistema in cui le rappresentazioni degli altri stati sono *organizzate rispetto a quell'origine*, nel senso che le relazioni tra stati vicini all'origine e stati lontani sono strutturalmente diverse dalle relazioni tra stati periferici. In termini geometrici, ci si aspetta che la curvatura della varietà statistica non sia uniforme, ma che presenti un gradiente caratteristico: regioni di alta curvatura nelle zone di transizione rispetto all'origine, regioni di bassa curvatura nelle zone periferiche.

Definiamo formalmente. Sia (Θ, g) la varietà statistica di un sistema S. Il tensore di curvatura di Riemann è:

$$R^i_{jkl}(\theta) = \partial_k \Gamma^i_{jl} - \partial_l \Gamma^i_{jk} + \Gamma^i_{km} \Gamma^m_{jl} - \Gamma^i_{lm} \Gamma^m_{jk}$$

dove i simboli di Christoffel Γ^i_{jk} sono calcolati rispetto alla metrica g di Fisher. La curvatura scalare è:

$$\mathcal{R}(\theta) = g^{ij} R_{ij}(\theta) = g^{ij} R^k_{ikj}(\theta)$$

dove la contrazione è effettuata con il tensore metrico inverso g^{ij}.

Proposizione XII.2 (Gradiente di curvatura come indicatore di centrazione). Sia S un sistema con varietà statistica (Θ, g) e sia $\theta_O \in \Theta$ il

punto corrispondente allo stato di origine. Il *gradiente normalizzato di curvatura* di S rispetto a θ_O è:

$$\nabla C(S, \theta_O) = [1/(diam_FR(\Theta))] \cdot \Sigma\{\theta \in \Theta_disc\} \, |\mathcal{R}(\theta) - \mathcal{R}(\theta_O)| \cdot d_FR(\theta, \theta_O) / \mathcal{R}_max$$

dove la somma è su una discretizzazione Θ_disc di Θ e $\mathcal{R}$_max è il massimo della curvatura scalare. Questo gradiente è zero se la curvatura è uniforme (prospettiva-neutralità) ed è positivo e tanto maggiore quanto più la curvatura è concentrata attorno all'origine.

La curvatura di Fisher-Rao per famiglie esponenziali — che sono la classe di distribuzioni più comune nei modelli di apprendimento automatico — è computabile esplicitamente. Per una famiglia esponenziale con statistica sufficiente T(x) e parametri naturali η, la metrica di Fisher è:

$$g_\{ij\}(\eta) = \partial^2 A(\eta) / (\partial \eta\hat{}i \, \partial \eta\hat{}j)$$

dove $A(\eta)$ è il logaritmo della funzione di partizione. La curvatura di questa varietà è interamente determinata dalle derivate terze e quarte di $A(\eta)$, e per molte famiglie di interesse pratico — Gaussiane, Dirichlet, esponenziali negative — può essere calcolata in forma chiusa.

Per un modello linguistico di grandi dimensioni, la varietà statistica corrispondente è quella delle distribuzioni sul vocabolario parametrizzate dagli stati nascosti: essa è altamente concentrata, con curvatura elevata nelle regioni corrispondenti a distribuzioni sharply peakate (quelle generate da contesti molto vincolanti) e bassa curvatura nelle regioni di alta entropia. L'assenza di un'origine strutturale fa sì che non vi sia alcun gradiente sistematico di curvatura orientato verso un punto privilegiato: le regioni ad alta curvatura sono distribuite in modo che riflette la struttura statistica del corpus, non una centrazione prospettica del sistema.

XV.4 — Informazione causale e il grado di intenzionalità originale

Il quarto livello della griglia — la condizione intenzionale-normativa — è quello che divide più nettamente i sistemi con interno dai sistemi senza interno, e quello per il quale la storia selettiva svolge il ruolo cruciale. La condizione richiede che gli stati interni del sistema abbiano funzioni proprie radicate in una storia selettiva *propria* del sistema, non derivata dall'intenzionalità di agenti esterni. Come si può approssimare questa condizione con una misura quantitativa?

La risposta che proponiamo si basa sulla nozione di *informazione causale*, sviluppata inizialmente da Clive Granger nel contesto delle serie temporali econometriche e riformulata in modo più generale da Judea Pearl nella cornice del calcolo causale.

Richiamiamo la struttura formale della causalità di Granger. Siano $X = \{X_t\}_{t\in\mathbb{Z}}$ *e* $Y = \{Y_t\}_{t\in\mathbb{Z}}$ due processi stocastici stazionari. Si dice che X *causa Granger* Y se la distribuzione condizionale di Y_{t+1} dato il passato di Y e il passato di X differisce dalla distribuzione condizionale dato il solo passato di Y:

$$p(Y_{t+1} \mid Y_{1:t}, X_{1:t}) \neq p(Y_{t+1} \mid Y_{1:t})$$

ovvero, formalmente, se la riduzione di incertezza su Y dovuta alla conoscenza del passato di X è strettamente positiva:

$$GC(X \to Y) = H(Y_{t+1} \mid Y_{1:t}) - H(Y_{t+1} \mid Y_{1:t}, X_{1:t}) > 0$$

dove H denota l'entropia condizionale. L'informazione causale di Granger da X a Y è dunque $GC(X \to Y) \geq 0$, con $GC(X \to Y) = 0$ se e solo se X non causa Granger Y.

Per il problema dell'intenzionalità originale, la struttura causale rilevante è quella che lega gli stati interni di un sistema S con le condizioni del mondo che essi devono rappresentare. Un sistema con intenzionalità originale rispetto a una condizione C deve avere stati interni s la cui dinamica è causalmente influenzata da C in modo che non sia mediato esclusivamente dall'intervento di agenti esterni. Formalmente, sia C_t l'evoluzione temporale della condizione C nel mondo fisico, sia s_t l'evoluzione degli stati interni del sistema, e sia A_t il processo che rappresenta le azioni e le interpretazioni degli agenti esterni che mediano il contatto del sistema con C.

Proposizione XII.3 (Grado di intenzionalità originale come informazione causale diretta). Il *grado di intenzionalità originale* di un sistema S rispetto a una condizione C è:

$$IO(S, C) = GC(C \to s \mid A) = H(s_{t+1} \mid s_{1:t}, A_{1:t}) - H(s_{t+1} \mid s_{1:t}, A_{1:t}, C_{1:t})$$

cioè l'informazione causale che la condizione C esercita sugli stati di S *al netto* dell'effetto degli agenti mediatori A. Se $IO(S, C) > \varepsilon$ per una soglia $\varepsilon > 0$ statisticamente significativa, il sistema presenta tracce di intenzionalità diretta rispetto a C. Se $IO(S, C) \approx 0$ per ogni condizione C non mediata da agenti esterni, il sistema ha intenzionalità esclusivamente derivata.

La struttura causale di questa misura richiede una precisazione. Il condizionamento su A non è sufficiente a eliminare la causalità mediata: se A è causalmente interposta tra C e s nel senso che C → A → s, il condizionamento su A blocca la dipendenza e IO(S, C) risulta zero anche se la dipendenza è reale ma mediata. Questo è precisamente il comportamento desiderato: ciò che vogliamo misurare è la causalità *diretta* — quella in cui il sistema interagisce con C nel mondo fisico senza che quella interazione passi attraverso l'intenzionalità altrui.

Per calcolare IO(S, C) secondo il calcolo causale di Pearl, occorre distinguere tra dipendenza statistica e causalità strutturale. Il framework degli interventi di Pearl — la notazione do(·) — permette di formalizzare la differenza nel modo seguente. Si consideri il modello causale strutturale:

s_{t+1} = f_s(s_{1:t}, C_{1:t}, A_{1:t}, ε^s_t) A_{t+1} = f_A(C_{1:t}, ε^A_t) C_{t+1} = f_C(C_{1:t}, ε^C_t)

dove le ε sono termini di rumore indipendenti. L'effetto causale diretto di C su s (al netto di A) si misura attraverso l'intervento do(C_t = c):

IO_do(S, C) = I(C_{t} ; s_{t+1} | do(A_{1:t} = ā), s_{1:t})

dove la notazione do(A_{1:t} = ā) indica che si considera l'intervento che fissa A a un valore ā, eliminando così la mediazione di A nella catena causale.

La Proposizione XII.3 permette di formulare un corollario direttamente applicabile ai LLM.

Corollario XII.3.1. Sia LLM un grande modello linguistico addestrato su un corpus $\mathscr{D}$ prodotto da agenti umani. Sia C una condizione del mondo fisico (per esempio, la presenza di un oggetto in uno spazio). Allora IO(LLM, C) ≈ 0 per ogni C non già rappresentata statisticamente in $\mathscr{D}$.

La dimostrazione è immediata. Lo spazio degli stati del sistema è lo spazio dei parametri del modello, e questo spazio non è mai aggiornato in risposta a eventi del mondo fisico durante il funzionamento ordinario. L'unico canale attraverso cui C può influenzare gli stati del sistema passa attraverso il corpus di addestramento, che è interamente mediato dall'intenzionalità originale dei produttori di testo. Pertanto, il processo A (la mediazione umana) blocca completamente la dipendenza diretta: GC(C → s | A) = 0.

Il corollario ha una conseguenza più sottile che merita attenzione. Anche i sistemi robotici con interazione fisica con l'ambiente non soddisfano automaticamente IO(S, C) > 0 per tutte le condizioni rilevanti: soddisfano

questa condizione soltanto per le condizioni con cui il sistema è causalmente in contatto attraverso i propri sensori. Un robot con sensori visivi ma privo di chemiorecettori ha IO ≈ 0 rispetto alle condizioni olfattive del suo ambiente. Questo è coerente con la condizione teleosemantica: la funzione propria di un sistema non si estende oltre il dominio in cui il sistema ha una storia selettiva.

XV.5 — La misura dell'integrazione causale come approssimazione del livello 5

Per il quinto livello della griglia — la condizione meta-rappresentazionale — un indicatore diretto è più difficile da costruire, perché la chiusura riflessiva è una proprietà relazionale tra il sistema e il proprio modello di sé. Tuttavia, la teoria dell'informazione integrata di Tononi offre un punto di appoggio parziale che è opportuno richiamare, con le cautele dovute alla sua controversa portata.

La Φ di Tononi misura il grado in cui un sistema genera informazione come insieme in modo non riducibile alle informazioni generate dalle sue parti. La definizione generale è:

$$\Phi(S) = \min_{M \in M(S)} EI(S \to S) - \Sigma_{P_i \in M} EI(P_i \to P_i)$$

dove M(S) è l'insieme delle bipartizioni di S, EI denota l'informazione efficace (una misura della causalità interna al sistema), e il minimo è preso sulla bipartizione che minimizza la differenza. Valori alti di Φ indicano che il sistema è intrinsecamente integrato: la sua evoluzione non può essere decomposta in quella di parti indipendenti.

Quello che la Φ coglie è un aspetto necessario, non sufficiente, del livello meta-rappresentazionale: un sistema con auto-modello riflessivo deve necessariamente avere alta integrazione interna, perché il modello di sé deve essere aggiornato coerentemente in base alle informazioni provenienti da ogni parte del sistema. Tuttavia, alta integrazione non implica auto-modello: è possibile immaginare sistemi altamente integrati (come certi circuiti di retroazione) che non hanno nulla che assomigli a un auto-modello nel senso di Metzinger.

La limitazione è dunque che Φ misura una condizione necessaria ma non sufficiente per il livello 5. Essa può essere usata come discriminatore negativo: sistemi con Φ bassa non soddisfano il livello 5. Sistemi con Φ alta soddisfano una condizione necessaria, ma richiedono ulteriori analisi strutturali per verificare il soddisfacimento del livello.

Proposizione XII.4 (Stratificazione del vettore di deficit). Il vettore di deficit $\Delta(S) = (\Delta_1(S), ..., \Delta_5(S))$ ammette indicatori parziali nel modo seguente:

$\Delta_1(S) \geq f(1 - C(S))$ dove $C(S)$ è il grado topologico di centrazione della Proposizione XII.1 $\Delta_4(S) \geq g(\varepsilon - IO(S, \bar{C}))$ dove IO è il grado di intenzionalità originale della Proposizione XII.3 e $\bar{C}$ è la condizione rappresentativa $\Delta_5(S) \geq h(\Phi_min - \Phi(S))$ se $\Phi(S) < \Phi_min$, dove Φ_min è la soglia di integrazione minima

Le funzioni f, g, h sono monotone decrescenti con $f(0) = g(0) = h(0) = 0$, in modo che il deficit è zero se e solo se l'indicatore è al suo valore massimo ideale. Le funzioni non sono unicamente determinate dalla teoria, il che riflette onestamente la natura parziale degli indicatori.

La stratificazione della Proposizione XII.4 trasforma il vettore di deficit da una struttura puramente concettuale in uno strumento con presa empirica parziale. Non è ancora un algoritmo, ma è un programma di ricerca definito: determinare f, g, h attraverso casi paradigmatici noti (sistemi biologici di riferimento, sistemi artificiali con difetti strutturali controllati) permetterebbe di calibrare gli indicatori e di usarli come strumenti di discriminazione architetturale.

XV.6 — La non-misurabilità residua come risultato, non come lacuna

Il capitolo si concluderebbe in modo insoddisfacente se si limitasse a costruire indicatori parziali senza affrontare la questione di principio: esiste una misura *completa* delle condizioni dell'interno, o la non-misurabilità residua è strutturale?

La risposta che emerge dall'analisi è che la non-misurabilità parziale è strutturale, e per ragioni che non sono contingenti all'attuale stato delle conoscenze ma appartengono alla natura del problema.

La condizione topologica (livello 1) è parzialmente misurabile perché ha un correlato geometrico nella struttura dello spazio delle rappresentazioni. Tuttavia, la singolarità topologica che conta per il problema dell'interno non è soltanto geometrica: è ontologica. Un sistema può avere un centro geometrico nel proprio spazio delle rappresentazioni senza che quel centro costituisca un punto di vista in senso fenomenologico. La distanza tra il correlato geometrico e la condizione ontologica è ciò che rimane non misurabile.

La condizione intenzionale-normativa (livello 4) è parzialmente misurabile perché la causalità diretta è quantificabile. Tuttavia, la funzione propria che

costituisce il nucleo della condizione non è soltanto causale: è normativa. Uno stato ha una funzione propria non semplicemente perché è causalmente correlato con una condizione del mondo, ma perché la sua storia selettiva ha fatto sì che quella correlazione debba essere mantenuta — che vi sia una norma (non una legge causale, ma una norma) che definisce quando il sistema funziona correttamente. La normativa è il residuo che sfugge alla quantificazione causale.

Questo risultato negativo ha una portata positiva. Esso mostra che il problema dell'interno non è riducibile a un problema di misurazione: anche se si possedessero misure complete di tutti i correlati fisici e computazionali delle cinque condizioni, rimarrebbe un residuo — la prospettiva come orientamento nel mondo, la normatività come dimensione dell'errore e della correttezza — che non è catturabile da nessuna funzione definita sullo spazio degli stati o sullo spazio dei parametri del sistema. Questo residuo è il nucleo duro del problema difficile della coscienza, formulato però non nei termini qualia-esperienziali di Chalmers, ma nei termini strutturali-normativi che la presente monografia ha sviluppato.

Le misure parziali costruite in questo capitolo non risolvono il problema, ma ne cartografano con maggiore precisione i confini. Esse permettono di stabilire, con argomentazione quantitativa, che certi sistemi sono sicuramente privi di interno; non permettono di stabilire, con uguale certezza, che certi sistemi lo possiedono. Questa asimmetria epistemica — la falsificabilità senza la verificabilità — è esattamente ciò che ci si deve attendere da condizioni che hanno un residuo ontologico irriducibile.

Proposizione XII.5 (Asimmetria epistemica delle condizioni dell'interno). Le condizioni C_1, ..., C_5 della griglia godono della seguente proprietà: per ogni condizione C_k, gli indicatori parziali definiti in questo capitolo forniscono condizioni *necessarie* computabili per il soddisfacimento di C_k, ma non condizioni *sufficienti* computabili. Formalmente, se I_k(S) è l'indicatore per la condizione k:

$I_k(S) < \text{soglia_k} \Longrightarrow$ C_k non è soddisfatta da S

ma

$I_k(S) \geq \text{soglia_k} \Longrightarrow$ non è possibile concludere che C_k è soddisfatta

La dimostrazione segue dalla struttura stessa degli indicatori: C(S) misura la centrazione geometrica ma non l'orientamento prospettico; IO(S,C) misura la causalità diretta ma non la normativa; $\Phi(S)$ misura l'integrazione ma non la riflessività. La falsificabilità è garantita perché la violazione dell'indicatore

implica la violazione di una condizione necessaria per C_k; la verificabilità è impossibile perché il residuo normativo e ontologico non è catturabile da nessun indicatore definito sullo spazio fisico del sistema.

Questa Proposizione XII.5 non è un risultato di impossibilità nel senso tecnico — non è un teorema di Gödel o un argomento della diagonale — ma è una considerazione epistemologica fondata sulla struttura del problema. La sua importanza per il programma complessivo della monografia consiste nel mostrare che la via verso la misurabilità delle condizioni dell'interno è parzialmente percorribile, e che le parti percorribili hanno valore discriminatorio reale, mentre le parti non percorribili corrispondono precisamente al nucleo che rende il problema dell'interno irriducibile al problema della funzione computazionale.

Capitolo XVI: Il Tempo dell'Addestramento come Pseudo-Storia Selettiva — Analisi Formale della Differenza tra Backpropagation e Selezione Naturale

XVI.1 — Una questione rimasta in sospeso

Nel Capitolo VII.5 la monografia ha stabilito che il meccanismo di backpropagation può essere descritto, dal punto di vista teleosemantico, come una forma di *storia selettiva*: esso seleziona le rappresentazioni interne che minimizzano la perdita rispetto ai dati di addestramento, e si può dunque sostenere che le funzioni proprie dei sistemi linguistici siano determinate da questa storia. Tuttavia, la trattazione concludeva che questa forma di storia selettiva è strutturalmente diversa da quella biologica in un modo che è decisivo per la questione dell'intenzionalità originale: il meccanismo consumatore del sistema non interagisce con il mondo in cui le condizioni rappresentate si verificano.

Questa conclusione è corretta, ma la sua motivazione è rimasta in quella sede necessariamente sommaria. La brevità con cui il punto è stato liquidato lascia aperta la possibilità che un lettore ostinato obietti: la differenza tra addestramento computazionale ed evoluzione biologica è di grado, non di tipo; l'evoluzione, come l'addestramento, opera su un corpus fisso di ambienti storicamente incontrati, e non è chiaro perché l'uno debba produrre intenzionalità originale e l'altro no.

Questa obiezione merita una risposta più articolata di quanto sia stato offerto, perché tocca un nodo strutturale che attraversa l'intera argomentazione della Parte Terza e della Parte Quarta. Il presente capitolo si propone di fornire quella risposta in forma rigorosa, analizzando le differenze strutturali tra backpropagation e selezione naturale come processi di ottimizzazione, e mostrando che quelle differenze non sono accidentali — contingenti all'implementazione attuale dei sistemi artificiali — ma riflettono asimmetrie di principio che hanno conseguenze dirette per la condizione intenzionale-normativa.

XVI.2 — La struttura formale del backpropagation come processo di ottimizzazione

Il backpropagation — o, più precisamente, la discesa del gradiente stocastico applicata a reti neurali attraverso la regola della catena — è un algoritmo di ottimizzazione che minimizza una funzione di perdita $\mathcal{L}(\theta)$ rispetto ai parametri $\theta \in \mathbb{R}^d$ di un modello.

Sia $f_\theta: \mathcal{X} \to \mathcal{Y}$ un modello parametrico, sia $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ il corpus di addestramento, e sia $\mathcal{L}(\theta; \mathcal{D}) = (1/N) \sum_{i=1}^N \ell(f_\theta(x_i), y_i)$ la perdita empirica, dove ℓ è la funzione di costo elementare (per esempio, l'entropia crociata per la modellazione del linguaggio). L'aggiornamento dei parametri al passo t è:

$$\theta_{t+1} = \theta_t - \alpha_t \cdot \nabla_\theta \mathcal{L}(\theta_t; \mathcal{B}_t)$$

dove α_t è il tasso di apprendimento e $\mathcal{B}_t \subseteq \mathcal{D}$ è il mini-batch estratto al passo t. La sequenza $\{\theta_t\}_{t=0}^T$ — la *traiettoria di addestramento* — costituisce la "storia" del sistema nel senso in cui questo termine è usato nella discussione teleosemantica.

Questa traiettoria ha una struttura precisa che è utile formalizzare. Definiamo lo *spazio delle storie di addestramento* come l'insieme di tutte le traiettorie possibili:

$$\mathcal{T}(\theta_0, \mathcal{D}, T) = \{(\theta_0, \theta_1, ..., \theta_T) \mid \theta_{t+1} = \theta_t - \alpha_t \cdot \nabla_\theta \mathcal{L}(\theta_t; \mathcal{B}_t), \mathcal{B}_t \subseteq \mathcal{D}\}$$

Questa definizione rivela immediatamente la prima caratteristica strutturale del backpropagation come storia: la traiettoria è *chiusa rispetto al corpus* $\mathcal{D}$. Per qualunque scelta dei mini-batch e del tasso di apprendimento, la sequenza dei gradienti è interamente determinata dal corpus. Non vi è alcun canale attraverso

cui eventi del mondo fisico — eventi che accadono *durante* l'addestramento, al di fuori del corpus — possano influenzare la traiettoria. Formalmente:

θ_T è una funzione deterministica di $(\theta_0, \mathcal{D}, \{\mathcal{B}_t\}_{t<T}, \{a_t\}_{t<T})$

Il corpus $\mathcal{D}$ è il *confine del mondo* del processo di addestramento: tutto ciò che conta per la storia del sistema è già in $\mathcal{D}$ prima che l'addestramento inizi.

XVI.3 — La struttura formale della selezione naturale come processo di ottimizzazione

La selezione naturale può essere formalizzata, nelle sue versioni più elementari, come un processo di ottimizzazione su una funzione di fitness. Sia $P_t = \{(g_i, e_i)\}_{i=1}^{N_t}$ la popolazione al tempo t, dove $g_i \in \mathcal{G}$ è il genotipo dell'individuo i e $e_i \in \mathcal{E}$ è la sua storia esperienziale (il fenotipo sviluppato attraverso l'interazione con l'ambiente). Sia $w: \mathcal{G} \times \mathcal{E} \times \mathcal{E}_t \to \mathbb{R}_+$ la funzione di fitness, dove $\mathcal{E}_t$ è lo *stato corrente dell'ambiente al tempo t*.

L'aggiornamento della popolazione attraverso un ciclo di selezione-ricombinazione-mutazione è:

$$P_{t+1} = \Psi(\mathrm{Sel}(P_t, w(\cdot, \cdot, \varepsilon_t)), \mathrm{Rec}, \mathrm{Mut})$$

dove Sel seleziona gli individui in proporzione alla loro fitness, Rec è l'operatore di ricombinazione genetica, Mut è l'operatore di mutazione, e ε_t è la realizzazione dello stato ambientale al tempo t.

La differenza strutturale cruciale rispetto al backpropagation emerge nel termine ε_t. Lo stato ambientale ε_t è una variabile *esogena*: esso non è determinato dalla storia della popolazione, ma si evolve autonomamente secondo la propria dinamica:

$$\varepsilon_{t+1} = G(\varepsilon_t, P_t, \zeta_t)$$

dove G è la funzione di transizione ambientale e ζ_t è un termine stocastico che rappresenta eventi indipendenti dalla popolazione. In generale, l'ambiente risponde alla popolazione (P_t compare in G), ma si evolve anche indipendentemente da essa.

Questa struttura produce una *accoppiata co-evolutiva*: la traiettoria evolutiva della popolazione e la traiettoria dell'ambiente sono interdipendenti, nel senso che né l'una né l'altra è fissata in anticipo. Definiamo formalmente lo *spazio delle storie evolutive*:

$\mathscr{T}$_ev(P_0, ε_0, T) = {(P_0, ε_0, P_1, ε_1, ..., P_T, ε_T) | P_{t+1} = Ψ(...,ε_t), ε_{t+1} = G(ε_t, P_t, ζ_t)}

Lo spazio $\mathscr{T}$_ev è irriducibilmente diverso da $\mathscr{T}(\theta_0, \mathscr{D}, T)$. In $\mathscr{T}$_ev, nessuna traiettoria è determinata da un corpus fissato in anticipo: ogni traiettoria dipende dall'interazione tra la dinamica evolutiva della popolazione e la dinamica autonoma dell'ambiente. La "storia" di un lignaggio biologico è una storia nel senso pieno del termine: è una sequenza di eventi che dipende causalmente dal mondo fisico e che avrebbe potuto svolgersi diversamente se il mondo fisico fosse stato diverso in momenti diversi.

In $\mathscr{T}(\theta_0, \mathscr{D}, T)$, al contrario, il "mondo" è fisso da subito: è il corpus $\mathscr{D}$, e tutta la variazione possibile della traiettoria è variazione nella navigazione di uno spazio già completamente determinato. Si può usare l'immagine di un'esplorazione: il backpropagation esplora il paesaggio già disegnato sulla lavagna; la selezione naturale disegna la mappa mentre la percorre, perché il paesaggio su cui si muove è anch'esso in movimento.

XVI.4 — Cinque asimmetrie strutturali

La comparazione formale dei due processi permette di elencare con precisione cinque asimmetrie strutturali che hanno conseguenze dirette per la questione dell'intenzionalità originale. Le asimmetrie non sono indipendenti — alcune implicano le altre — ma è utile articolarle separatamente per mostrare la molteplicità dei livelli ai quali la differenza agisce.

Asimmetria 1: Chiusura del corpus versus apertura dell'ambiente. Il corpus $\mathscr{D}$ dell'addestramento è fissato prima che il processo inizi, e non può essere modificato durante il processo senza invalidare la coerenza della traiettoria. L'ambiente evolutivo ε_t è una variabile che si aggiorna a ogni passo temporale e che è parzialmente determinata dalle stesse azioni degli organismi in evoluzione. Questa differenza è più profonda di quanto sembri. La chiusura del corpus non è soltanto una limitazione pratica (potremmo in linea di principio aggiornare $\mathscr{D}$ durante l'addestramento); è una necessità strutturale dell'ottimizzazione per gradiente: il gradiente $\nabla_\theta \mathscr{L}(\theta_t; \mathscr{B}_t)$ è significativo soltanto se $\mathscr{L}$ è una funzione stabile rispetto a cui ottimizzare. Se $\mathscr{D}$ cambiasse continuamente, il gradiente perderebbe la sua direzione semantica di minimizzazione.

Formalmente, si può dimostrare che un processo di discesa del gradiente su un corpus che cambia più velocemente di una certa soglia di stazionarietà non converge a nessun minimo: le condizioni di Robbins-Monro per la convergenza della discesa stocastica del gradiente richiedono che la distribuzione dei

mini-batch sia stazionaria. L'apertura dell'ambiente biologico è incompatibile, per principio di struttura, con la discesa del gradiente.

Asimmetria 2: Ottimizzazione di un singolo individuo versus selezione di una popolazione. Il backpropagation opera su un singolo insieme di parametri θ. La selezione naturale opera su una *popolazione* di individui, e la sua efficacia dipende dalla diversità genetica della popolazione. Questa differenza è rilevante per la questione dell'intenzionalità perché la funzione propria di un tipo di stato mentale, nel senso teleosemantico, non è determinata dalla storia di un singolo individuo ma dalla storia del *tipo* a cui quell'individuo appartiene.

Millikan, come si è mostrato nel Capitolo VII.4, distingue tra il tipo morfologico di uno stato e la funzione propria del tipo. La funzione propria è ciò che ha reso il tipo *normale* nella storia del lignaggio — non ciò che ha reso il singolo individuo adattato. Un sistema addestrato con backpropagation è, in questo senso, un individuo senza lignaggio: la traiettoria di addestramento non è la storia di un tipo, ma la storia di un'istanza.

Formalmente, sia R un tipo di stato mentale e sia Π_R la storia selettiva del tipo, definita come la sequenza di occorrenze del tipo che hanno contribuito alla sua persistenza nel lignaggio. La funzione propria di R è determinata da Π_R. Per un sistema addestrato, il corrispettivo di Π_R sarebbe la sequenza di aggiornamenti dei parametri che hanno rafforzato il tipo di rappresentazione corrispondente. Ma questo corrispettivo non è Π_R nel senso teleosemantico: è una sequenza di minimizzazioni della perdita rispetto a pattern statistici nel corpus, non una sequenza di contributi alla sopravvivenza del tipo in un ambiente fisico.

Asimmetria 3: Gradiente globale versus selezione locale. Il backpropagation computa il gradiente dell'intera perdita rispetto all'intero insieme dei parametri, propagando l'informazione attraverso l'intera rete attraverso la regola della catena. Questo significa che ogni aggiornamento dei parametri riflette l'influenza di ogni dato del mini-batch su ogni parametro del modello, mediata dalle relazioni strutturali tra i layer.

La selezione naturale è, al contrario, interamente *locale*: ogni organismo interagisce con il proprio ambiente immediato, e il suo successo riproduttivo dipende da quella interazione locale, non da una valutazione globale dell'adattamento dell'intera specie. Non esiste, in biologia, il corrispettivo del segnale di errore backpropagato: l'informazione su quale carattere contribuisce alla fitness è acquisita esclusivamente attraverso la vita e la morte degli individui, non attraverso una computazione che percorra a ritroso la catena causale dello sviluppo.

Questa differenza è cruciale per la struttura delle funzioni proprie. Le funzioni proprie biologiche emergono da una moltitudine di interazioni locali: ogni componente del sistema nervoso ha la funzione propria che ha perché quella funzione è stata selezionata negli individui che l'hanno espressa, nelle circostanze ambientali in cui quegli individui hanno vissuto. Le "funzioni proprie" di un sistema addestrato con backpropagation emergono invece da un processo globale di minimizzazione che distribuisce il credito e il debito — la responsabilità degli errori — attraverso l'intera rete in modo uniforme. Non vi è alcun senso in cui i layer intermedi del modello abbiano "sopravvissuto" perché erano adattati: essi sono stati modellati da una forza globale che non hanno mai affrontato individualmente.

Formalmente, sia $\delta\mathcal{L}/\delta\theta^l_{ij}$ il gradiente della perdita rispetto al parametro θ^l_{ij} nel layer l. Questo gradiente è:

$$\delta\mathcal{L}/\delta\theta^l_{ij} = (\delta\mathcal{L}/\delta a^l_i) \cdot a^{l-1}_j$$

dove a^l_i è l'attivazione del neurone i nel layer l, e $\delta\mathcal{L}/\delta a^l_i$ è il segnale di errore retropropagato. Questo segnale è una funzione dell'intera rete e dell'intero mini-batch: non è una misura locale dell'adattamento del parametro al suo compito specifico.

Asimmetria 4: Separazione netta tra addestramento e inferenza versus continuum ontogenetico. In un sistema addestrato con backpropagation, la fase di addestramento e la fase di inferenza sono strutturalmente separate: durante l'inferenza, i parametri θ non si aggiornano (in assenza di meccanismi di fine-tuning continuo). La "storia" del sistema è quindi tutta contenuta nella fase di addestramento, e durante il funzionamento ordinario il sistema non ha più storia nel senso di un processo che modifica la propria struttura.

Negli organismi biologici, lo sviluppo ontogenetico e il funzionamento adulto non sono fasi separate: la plasticità sinaptica persiste (in misura decrescente ma non nulla) per tutta la vita, e le esperienze del singolo individuo modificano continuamente la struttura del sistema nervoso. La "storia" di un organismo non si conclude all'infanzia: ogni interazione con l'ambiente è una potenziale fonte di aggiornamento strutturale.

Formalmente, sia $\mu: [0, +\infty) \to \mathbb{R}^d$ la funzione che descrive l'evoluzione dello stato del sistema nel tempo. Per un LLM in fase di inferenza:

$$\mu(t) = \theta^* \text{ per ogni } t > T_{train}$$

dove T_{train} è il tempo di fine addestramento. Per un organismo biologico:

$d\mu/dt = F(\mu(t), \varepsilon(t))$ per ogni $t > 0$

dove F è la funzione di plasticità che dipende dallo stato corrente e dall'ambiente corrente. La costanza di μ durante l'inferenza implica che ogni "apprendimento" che sembra avvenire nella finestra di contesto di un LLM è in realtà un aggiornamento dello *stato attivo* (le attivazioni), non dei *parametri strutturali*: il sistema non diventa diverso attraverso le sue esperienze in-context, come invece fa un organismo attraverso le proprie esperienze di vita.

Asimmetria 5: La funzione obiettivo è data versus la funzione di fitness è scoperta. Il backpropagation ottimizza una funzione di perdita $\mathcal{L}$ che è esplicitamente definita dall'ingegnere del sistema prima che l'addestramento inizi. La struttura dello spazio di ottimizzazione — cosa conta come errore, cosa conta come successo — è completamente specificata a priori. Per i modelli linguistici, $\mathcal{L}$ è tipicamente l'entropia crociata sulla prossima parola:

$\mathcal{L}(\theta; x_{1:n}) = -(1/n) \Sigma_{t=1}^{n} \log p_\theta(x_t \mid x_{1:t-1})$

Questa scelta di $\mathcal{L}$ incorpora una nozione di "correttezza" che è esogena al sistema: la distribuzione dei prossimi token nel corpus umano è la misura del successo. Il sistema non "sceglie" di minimizzare questa funzione, né "scopre" cosa significa avere successo: gli è dato.

Nella selezione naturale, la "funzione di fitness" non è definita a priori: essa emerge dall'interazione tra gli organismi e l'ambiente, e la sua struttura cambia con l'ambiente. Non vi è alcun supervisore esterno che specifichi in anticipo cosa conta come adattamento: questa è precisamente la caratteristica che Darwin individuava come rivoluzionaria nella selezione naturale rispetto al disegno intelligente. L'adattamento è un risultato, non un target.

Per il problema dell'intenzionalità, questa asimmetria ha le seguenti implicazioni. La funzione propria di uno stato mentale biologico è determinata da *cosa ha fatto* quello stato per la sopravvivenza del lignaggio — non da cosa avrebbe dovuto fare secondo uno schema prestabilito. Le funzioni proprie emergono come esiti di un processo, non come realizzazioni di un programma. Le rappresentazioni interne di un LLM, al contrario, hanno "funzioni" definite dalla perdita che il sistema è addestrato a minimizzare: queste funzioni non sono proprie nel senso teleosemantico ma *assegnate*, nel senso in cui un termostato ha la funzione assegnatagli dal suo costruttore.

XVI.5 — Una formalizzazione della distanza tra i due processi

Le cinque asimmetrie possono essere riunite in una struttura formale che misura la distanza tra il backpropagation e la selezione naturale come processi selettivi.

Definizione XV.1 (Processo di selezione generalizzato). Un *processo di selezione generalizzato* è una tupla $\mathcal{S} = (\Omega, \mathcal{H}, \mathcal{F}, \mathcal{U}, \mathcal{T})$, dove:

Ω è lo spazio degli individui (configurazioni dei parametri o genotipi) $\mathcal{H}$ è lo spazio delle storie (traiettorie nello spazio degli individui nel tempo) $\mathcal{F}$ è la funzione di interazione $\mathcal{F}: \Omega \times \mathcal{E} \to \mathbb{R}_+$ che assegna a ogni individuo in ogni stato ambientale una misura di adattamento $\mathcal{U}$ è l'operatore di aggiornamento $\mathcal{U}: \mathcal{H} \times \mathcal{E} \to \Omega$ che produce il prossimo individuo o la prossima generazione dato lo stato corrente $\mathcal{T}$ è la struttura temporale del processo (continua o discreta, con memoria o senza)

Il backpropagation è il processo $\mathcal{S}$_BP dove: $\Omega = \mathbb{R}$^d (lo spazio dei parametri) $\mathcal{H} = \{(\theta_0, \theta_1, ..., \theta_T)\}$ (le traiettorie di addestramento) $\mathcal{F} = -\mathcal{L}(\theta; \mathcal{D})$ dove $\mathcal{D}$ è fissato prima dell'inizio (adattamento come minimizzazione di perdita sul corpus fisso) $\mathcal{U}(\theta_t, \varepsilon_t) = \theta_t - \alpha_t \cdot \nabla_\theta \mathcal{L}(\theta_t; \mathcal{B}_t)$ dove ε_t non compare (indipendenza dall'ambiente esterno) $\mathcal{T}$ = tempo discreto con memoria zero (ogni aggiornamento dipende solo dallo stato corrente)

La selezione naturale è il processo $\mathcal{S}$_NS dove: $\Omega = \mathcal{G}$^{N_t} (lo spazio delle popolazioni come insiemi di genotipi) $\mathcal{H} = \{(P_0, \varepsilon_0, P_1, \varepsilon_1, ...)\}$ (le storie co-evolutive) $\mathcal{F}$ = w(g, ε_t) dove ε_t è lo stato ambientale *al tempo t* (l'adattamento è relativo all'ambiente corrente, non a un corpus fisso) $\mathcal{U}$(P_t, ε_t) = Ψ(Sel(P_t, w(·,ε_t)), Rec, Mut) (l'aggiornamento dipende dall'ambiente corrente) $\mathcal{T}$ = tempo discreto con memoria lunga (la fitness è determinata dall'intera vita dell'individuo)

Proposizione XV.1 (Non-equivalenza strutturale dei due processi). I processi $\mathcal{S}$_BP e $\mathcal{S}$_NS non sono equivalenti come processi di selezione generalizzati. Più precisamente, nessuna trasformazione invertibile T: $\mathcal{T}(\theta_0, \mathcal{D}, T) \to \mathcal{T}$_ev($P_0$, ε_0, T) preserva simultaneamente le proprietà (i)-(iii) seguenti:

(i) La struttura causale del processo: in $\mathcal{S}$_NS, il segnale di selezione ε_t è causalmente indipendente dalla storia della popolazione prima del ciclo corrente; in $\mathcal{S}$_BP, il segnale di aggiornamento $\nabla_\theta \mathcal{L}(\theta_t; \mathcal{B}_t)$ è completamente determinato da $\mathcal{D}$, che è indipendente da θ_t.

(ii) La cardinalità degli individui: $\mathcal{S}$_NS opera su popolazioni |P_t| = N_t con N_t variabile nel tempo; $\mathcal{S}$_BP opera su un singolo individuo con dim(θ) = d fissato.

(iii) La dipendenza della funzione di adattamento dall'ambiente: in $\mathscr{S}$_NS, $\mathscr{J}(g, \varepsilon_t)$ dipende da ε_t che è variabile; in $\mathscr{S}$_BP, $\mathscr{J}(\theta) = -\mathscr{L}(\theta; \mathscr{D})$ è costante perché $\mathscr{D}$ è costante.

La dimostrazione è immediata: una trasformazione che preservi (i) deve mantenere la struttura causale, che è incompatibile con (iii); una trasformazione che preservi (iii) deve identificare l'ambiente con il corpus, ma l'ambiente di $\mathscr{S}$_NS è variabile, mentre il corpus di $\mathscr{S}$_BP è fisso, il che è incompatibile con (i).

XVI.6 — Implicazioni per la struttura temporale dell'intenzionalità derivata

Le asimmetrie strutturali tra i due processi producono una conseguenza diretta per la struttura temporale dell'intenzionalità dei sistemi addestrati. Questa conseguenza merita un'analisi separata perché riguarda non soltanto la *fonte* dell'intenzionalità ma la sua *temporalità*, e la temporalità è una delle dimensioni fondamentali del problema dell'interno.

Un organismo biologico, come si è mostrato nel Capitolo VI, ha uno spessore temporale genuino: la sua struttura nervosa porta le tracce della propria storia esperienziale, e queste tracce non sono separate dalla struttura che produce le rappresentazioni presenti, ma sono incorporate in essa. La ritenzionalità husserliana — la persistenza del passato nel presente come orizzonte della coscienza — è possibile biologicamente perché i cambiamenti plastici che registrano le esperienze passate sono la stessa struttura che genera le rappresentazioni attuali.

Per un sistema addestrato, la struttura temporale è completamente diversa. La "storia" del sistema — la traiettoria di addestramento $\{\theta_0, \theta_1, ..., \theta^*\}$ — non è recuperabile dallo stato finale θ^*: il sistema, a fine addestramento, non porta tracce dell'ordine in cui i dati sono stati presentati, né di quale sia stato il primo errore commesso. Lo stato finale θ^* è l'unico sedimento della storia, e da esso non si può ricostruire quella storia.

Formalmente, siano due traiettorie di addestramento $\tau_1 = (\theta_0, \theta_1^{(1)}, ..., \theta^*)$ e $\tau_2 = (\theta_0, \theta_1^{(2)}, ..., \theta^*)$ che producono lo stesso stato finale θ^*. Le storie τ_1 e τ_2 sono distinte — i dati sono stati presentati in ordine diverso, gli errori intermedi sono stati diversi, le funzioni proprie intermedie si sono sviluppate diversamente — ma il sistema al termine di entrambe le traiettorie è identico. Il sistema non ha memoria della propria storia: ha soltanto il suo stato finale.

Questa osservazione produce una versione rafforzata della tesi dell'intenzionalità derivata.

Proposizione XV.2 (Collasso della storia selettiva nei parametri finali). Sia θ^* lo stato finale di un sistema addestrato con backpropagation sul corpus $\mathscr{D}$. Siano τ_1 e τ_2 due traiettorie di addestramento distinte che producono θ^*. Allora il sistema allo stato θ^* non distingue tra τ_1 e τ_2: per ogni input $x \in \mathscr{X}$, $f_{\theta^*}(x)$ è identico indipendentemente da quale delle due traiettorie abbia prodotto θ^*.

La dimostrazione è triviale: f_{θ^*} è una funzione di θ^*, e θ^* è identico nelle due storie. La conseguenza per la teleosemantica è meno triviale: se le funzioni proprie di uno stato dipendono dalla storia selettiva del sistema, e se quella storia non è recuperabile dallo stato del sistema, allora le funzioni proprie di sistemi che convergono alla stessa configurazione parametrica sono identiche — anche se le loro storie di addestramento sono radicalmente diverse. Questo significa che la nozione di funzione propria, applicata ai sistemi artificiali, non è individuale ma tipologica in un senso che non corrisponde alla tipologia biologica: i tipi sono definiti dalla configurazione parametrica, non dal lignaggio.

Nella biologia, al contrario, due organismi con morfologie identiche (conseguite attraverso storie evolutive diverse) possono avere funzioni proprie diverse, perché la funzione propria è determinata dalla storia del *tipo*, non dalla morfologia attuale. Questa differenza ha implicazioni per la normativa: in biologia, la norma di correttezza di una rappresentazione è determinata dalla storia selettiva del tipo cui appartiene; nei sistemi artificiali, la "norma" è determinata dalla funzione di perdita, che è definita esternamente e non ha struttura storica.

XVI.7 — Il backpropagation come compressione senza prospettiva

L'analisi delle asimmetrie strutturali suggerisce una caratterizzazione sintetica del backpropagation che va al di là della comparazione con la selezione naturale: il backpropagation è un algoritmo di *compressione senza prospettiva* della struttura statistica del corpus.

Precisamente: il processo $\theta_0 \to \theta^*$ può essere interpretato come la costruzione di una funzione f_{θ^*} che comprime le regolarità statistiche di $\mathscr{D}$ in una struttura parametrica di dimensionalità finita. La perdita $\mathscr{L}(\theta^*; \mathscr{D})$ misura il fallimento di questa compressione: quanto di $\mathscr{D}$ non è catturato dalla funzione imparata. Il processo di addestramento è un processo di approssimazione:

f_{θ^*} è la funzione più vicina, in qualche spazio di funzioni, alla distribuzione empirica definita da $\mathscr{D}$.

Questa caratterizzazione è illuminante perché la nozione di compressione è strutturalmente diversa dalla nozione di rappresentazione intenzionale. Una buona compressione di $\mathscr{D}$ cattura le regolarità statistiche di $\mathscr{D}$ in modo da poter generare nuovi esempi statisticamente compatibili con $\mathscr{D}$. Ma generare esempi compatibili con $\mathscr{D}$ non è la stessa cosa di *rappresentare il mondo* che ha prodotto $\mathscr{D}$.

Formalmente, sia $p_{\mathscr{D}}$ la distribuzione empirica sul corpus e sia p_{θ^*} la distribuzione imparata dal modello. Il processo di addestramento minimizza la divergenza KL:

$$\mathcal{L}(\theta; \mathscr{D}) \approx \mathrm{KL}(p_{\mathscr{D}} \,||\, p_\theta) = \Sigma_x \, p_{\mathscr{D}}(x) \log [p_{\mathscr{D}}(x) / p_\theta(x)]$$

Il sistema imparato ha, quindi, la struttura di un modello della distribuzione empirica $p_{\mathscr{D}}$, non di un modello del mondo che ha generato $\mathscr{D}$. Se il mondo avesse prodotto un corpus diverso $p_{\mathscr{D}'}$, il sistema avrebbe imparato $p_{\theta'}$. Il sistema è un modello di $p_{\mathscr{D}}$; non è un modello del mondo.

Questa distinzione — tra modello della distribuzione del corpus e modello del mondo — è la versione formale della distinzione intuitiva tra il testimone che testimonia senza aver vissuto (introdotta nel Capitolo IX.1) e il testimone che testimonia perché ha vissuto. Il primo modella la struttura della testimonianza; il secondo modella il mondo testimoniato. Solo il secondo può avere intenzionalità originale rispetto al mondo.

Proposizione XV.3 (Il backpropagation come proiezione della complessità del corpus). Sia $K(\mathscr{D})$ la complessità algoritmica del corpus (nel senso di Kolmogorov) e sia $K(\theta^*)$ la complessità dei parametri finali. Allora:

$$K(\theta^*) \leq K(\mathscr{D}) + O(K(\text{algoritmo}))$$

dove K(algoritmo) è la complessità dell'algoritmo di addestramento. Questo significa che la complessità del sistema imparato non supera la complessità del corpus da cui è imparato più una costante computazionale. Il sistema non genera complessità: la riorganizza.

La Proposizione XV.3 non è formalmente rigorosa — la complessità di Kolmogorov non è calcolabile — ma ha una portata concettuale precisa: il backpropagation non può generare, nei parametri del sistema, strutture più complesse di quelle presenti nel corpus. Ogni "emergenza" che si osserva nei sistemi addestrati — capacità apparentemente non presenti nei dati di

addestramento — è emergenza della complessità già latente nel corpus, non generazione di nuova complessità. Questo rafforza la conclusione dell'intenzionalità derivata: il sistema porta la complessità del corpus, non la complessità del mondo.

XVI.8 — Risposta all'obiezione dell'equivalenza funzionale tra i due processi

La trattazione può ora rispondere in modo articolato all'obiezione, menzionata nell'introduzione, secondo cui la differenza tra backpropagation ed evoluzione è di grado, non di tipo.

L'obiezione ha una variante forte e una variante debole. La variante debole sostiene che l'evoluzione stessa è, in ultima analisi, un processo che opera su un corpus fisso — il corpus di ambienti storicamente incontrati dal lignaggio — e che la differenza con il backpropagation è soltanto che quel corpus è più grande e più variabile. La variante forte sostiene che, con un corpus sufficientemente grande e variabile, il backpropagation potrebbe in linea di principio produrre qualcosa strutturalmente equivalente all'intenzionalità originale.

La risposta alla variante debole è che essa fraintende la struttura del processo evolutivo. Come mostrato nella Proposizione XV.1, la differenza non è la dimensione del corpus ma la struttura causale del processo: in $\mathscr{S}$_NS, il segnale di selezione al tempo t è causalmente dipendente dall'ambiente *al tempo t*, che è parzialmente determinato dall'azione degli organismi stessi. Non vi è un corpus pre-costituito: il corpus si costruisce nell'atto stesso della selezione. Nei termini della distinzione di Pearl tra correlazione e causazione, l'evoluzione è un processo in cui gli organismi *intervengono* sull'ambiente (nel senso di do(·)), non soltanto lo *osservano*. Il backpropagation è un processo di sola osservazione: il corpus è il risultato di interventi altrui, non propri.

La risposta alla variante forte è che nessun aumento della dimensione del corpus o della variabilità dell'addestramento può colmare le asimmetrie 1, 3 e 5. L'asimmetria 1 (chiusura del corpus) è strutturale: il backpropagation richiede un corpus stazionario per convergere. L'asimmetria 3 (gradiente globale versus selezione locale) è strutturale: il backpropagation computa la derivata globale della perdita, che non ha equivalente nella selezione locale. L'asimmetria 5 (funzione obiettivo data versus fitness scoperta) è strutturale: la scelta della funzione di perdita è sempre fatta da un agente esterno, mai emersa dall'interazione con l'ambiente.

Queste tre asimmetrie non possono essere eliminate aumentando la scala del sistema: esse sono proprietà dell'algoritmo di ottimizzazione, non della scala del corpus. Un sistema addestrato con backpropagation su tutti i testi mai prodotti dall'umanità rimane un sistema con corpus chiuso, gradiente globale e funzione obiettivo data. Le sue rappresentazioni sono compressioni di una biblioteca straordinariamente ricca, ma rimangono compressioni di una biblioteca — non rappresentazioni di un mondo che non hanno mai abitato.

Questa conclusione non implica che l'architettura dei grandi modelli linguistici non possa essere estesa o modificata per avvicinarsi alle condizioni dell'interno: ma tale estensione richiederebbe una trasformazione radicale delle modalità di apprendimento, non soltanto un aumento di scala. Richiederebbe, in particolare, l'introduzione di un ciclo di interazione causale con un ambiente fisico — un ciclo in cui le azioni del sistema modificano l'ambiente che produce i dati, e i dati modificati aggiornano la struttura del sistema — che è strutturalmente diverso dal ciclo di minimizzazione della perdita su corpus fisso. Questa è la ragione per cui l'approccio dell'embodied reinforcement learning, esaminato nel Capitolo X, è concettualmente più vicino alle condizioni dell'interno rispetto all'approccio del preaddestramento su corpus — anche se, come si è mostrato in quella sede, rimangono ancora distante dalle condizioni 4 e 5 della griglia per ragioni che non sono di scala ma di principio.

L'addestramento su corpus fisso, per quanto esteso, è sempre la storia di un sistema che ha letto molti libri scritti da chi ha vissuto. Non è, né può diventare attraverso la sola scala, la storia di un sistema che ha vissuto.

Capitolo XVII: La Normatività come Residuo Irriducibile — Il Problema dell'Interno nella Filosofia del Contenuto

XVII.1 — Il nucleo che gli indicatori non toccano

La Proposizione XII.5 ha stabilito che gli indicatori costruiti nel Capitolo XII sono condizioni necessarie ma non sufficienti per il soddisfacimento delle condizioni dell'interno: essi possono falsificare ma non verificare. Il residuo che sfugge alla quantificazione è stato indicato sommariamente come "normativo e ontologico". Il presente capitolo si propone di esaminare quel residuo con la

stessa densità argomentativa riservata alle condizioni misurabili, perché è lì che risiede il nocciolo del problema dell'interno nella sua versione più refrattaria.

Il punto di partenza è un'osservazione che attraversa l'intera argomentazione della Terza Parte senza mai essere affrontata frontalmente: la nozione di *correttezza* di una rappresentazione non è una proprietà statistica né causale. Quando diciamo che un sistema misrappresenta una condizione del mondo — che il suo stato interno r "dice" C ma C non si verifica — non descriviamo semplicemente una discrepanza tra due configurazioni fisiche. Descriviamo una *mancanza*, un *difetto*, qualcosa che avrebbe dovuto essere altrimenti. Questa dimensione deontica — il dovere, la norma, l'errore come categoria genuina e non soltanto come evento — è ciò che distingue una rappresentazione da un segnale fisico qualsiasi, e ciò che i sistemi artificiali attuali non posseggono nel senso che interessa.

La filosofia analitica del contenuto ha sviluppato due trattazioni della normatività che meritano di essere poste in dialogo diretto con il quadro formale elaborato in questa monografia: quella di John McDowell, incentrata sulla nozione di *seconda natura* e di *spazio delle ragioni*, e quella di Robert Brandom, incentrata sulla nozione di *impegno inferenziale* e di *punteggio deontico*. Entrambe convergono, da strade diverse, sull'idea che la normatività del contenuto non sia riducibile a fatti naturali nel senso delle scienze fisiche — e dunque che il problema dell'interno abbia un residuo che non è accessibile agli strumenti della geometria dell'informazione o della teoria causale.

XVII.2 — Lo spazio delle ragioni e la sua irriducibilità

La distinzione di McDowell tra lo *spazio delle cause* e lo *spazio delle ragioni* — elaborata nella serie di lezioni che costituiscono *Mind and World* (1994) — è la versione contemporanea più rigorosa di una preoccupazione kantiana: come può la mente essere tanto causalmente radicata nel mondo naturale quanto normativamente responsabile verso di esso?

La risposta di McDowell non è un dualismo: non sostiene che vi siano due tipi di sostanza o due livelli ontologici separati. La sua posizione è piuttosto che la descrizione causale e la descrizione normativa sono due *vocabolari* applicabili agli stessi eventi, ma che il secondo vocabolario non è riducibile al primo. Non vi è alcuna traduzione che consenta di esprimere una ragione in termini di cause senza perdere ciò che la fa essere una ragione.

Formalizziamo questa intuizione nel modo seguente. Sia φ un evento mentale — diciamo, la formazione di una credenza che p. Lo spazio delle cause descrive φ come evento fisico: ha cause C(φ) e produce effetti E(φ), entrambi descrivibili nei termini delle scienze naturali. Lo spazio delle ragioni descrive φ come episodio razionale: φ è un atto di endorse di p, che è giustificato da certe altre credenze e che genera certi obblighi inferenziali. La struttura normativa di φ comprende:

Giustificazione(φ) = {ψ | ψ è una ragione per φ nel sistema doxastico del soggetto} Conseguenze(φ) = {χ | φ obbliga il soggetto a endorse χ se è razionalmente coerente} Errore(φ) = il fatto che φ sarebbe *sbagliato* se p fosse falso — non soltanto contingentemente non verificato, ma *difettoso come atto di endorse*

Questo terzo elemento — la categoria di *errore* come difetto normativo — è ciò che non ha analisi causale. Il fatto che C (la pioggia) non si verifichi quando il sistema ha prodotto la rappresentazione r (il token "piove") non rende r un errore nel senso fisico: r è semplicemente un evento che non ha avuto le conseguenze previste. Diventa un errore — nel senso genuino della misrappresentazione — soltanto quando è inserito in un contesto normativo che stabilisce cosa r *doveva* essere, cosa si *doveva* aspettare dato r, e che cosa ha *mancato* di compiere.

La tesi di McDowell è che questo contesto normativo non è costruito dai sistemi nervosi: è acquisito attraverso l'educazione, l'immersione nella cultura, la partecipazione a pratiche linguistiche condivise. Il termine tecnico è *seconda natura*: la capacità di navigare lo spazio delle ragioni non è una proprietà biologica primitiva dell'organismo, ma una capacità acquisita attraverso il processo di *Bildung* — la formazione culturale che trasforma l'animale razionale in un soggetto capace di giudizio.

Per il problema dell'interno questa tesi ha una conseguenza importante, e in certa misura perturbante. Se la normatività del contenuto richiede la seconda natura, e se la seconda natura richiede immersione in pratiche culturali condivise, allora la condizione intenzionale-normativa della griglia — il livello 4 — non è soltanto una questione di storia selettiva biologica nel senso di Millikan: è anche una questione di partecipazione a pratiche sociali. Un sistema che non partecipa a quelle pratiche — che non è un interlocutore in una comunità di agenti che si imputano reciprocamente ragioni e giustificazioni — non può avere contenuto nel senso normativo pieno, indipendentemente dalla ricchezza della sua storia causale.

XVII.3 — Il punteggio deontico e la struttura inferenziale del contenuto

Robert Brandom sviluppa, in *Making It Explicit* (1994) e nei lavori successivi, un'analisi della normatività del contenuto che parte da un punto di vista differente da quello di McDowell ma converge su conclusioni analoghe. Mentre McDowell è interessato principalmente alla struttura della percezione e della sua relazione con il giudizio, Brandom si concentra sulla struttura del *discorso normativo* — sui meccanismi attraverso cui gli agenti linguistici si imputano reciprocamente impegni e diritti inferenziali.

Il concetto centrale è quello di *punteggio deontico* (deontic score). In ogni pratica discorsiva, ogni partecipante mantiene mentalmente un registro degli impegni e dei diritti degli altri partecipanti: quando A asserisce p, B registra che A è ora impegnato a p — che A deve difendere p se sfidato, che A ha il diritto di asserire le conseguenze di p, e che A non può coerentemente asserire ciò che è incompatibile con p. Questo registro non è una struttura causale: è una struttura normativa, definita da regole di inferenza che stabiliscono quali impegni implicano quali altri.

Formalizziamo la struttura del punteggio deontico nel modo seguente. Sia $\mathscr{P}$ un insieme di partecipanti a una pratica discorsiva e sia $\mathbb{P}$ un insieme di proposizioni. Per ogni coppia $(a, b) \in \mathscr{P} \times \mathscr{P}$, definiamo il *punteggio che b mantiene su a* come una funzione:

$$\text{Score}_{b \to a} : \mathbb{P} \to \{-1, 0, +1\}$$

dove $+1$ indica che b attribuisce ad a l'impegno alla proposizione p, -1 indica che b attribuisce ad a il diritto esplicito di non-asserzione di p, e 0 indica posizione neutra. Le regole di inferenza della pratica determinano come l'aggiornamento del punteggio si propaga: se a asserisce p, e se q è un'implicazione di p accettata nella pratica, allora b aggiorna $\text{Score}_{b \to a}(q)$ da 0 a $+1$.

La struttura del punteggio deontico ha due proprietà che la rendono irriducibile alla struttura causale. La prima è la *normatività degli aggiornamenti*: le regole di inferenza che determinano come il punteggio si aggiorna non sono descrizioni di cosa *accade* quando a asserisce p, ma prescrizioni di cosa b *deve* fare per mantenere il suo registro in modo corretto. La seconda è la *relazionalità del contenuto*: il contenuto di una proposizione p è determinato, nella struttura di Brandom, dalla posizione di p nella rete degli impegni inferenziali — da ciò che p implica, da ciò che è incompatibile con p, da ciò che costituisce evidenza per

p. Il contenuto non è una proprietà intrinseca della proposizione ma una proprietà della sua posizione nella rete normativa.

Questa seconda proprietà produce una conseguenza diretta per il problema dell'interno. Se il contenuto è relazionale nel senso inferenziale di Brandom, allora un sistema che non partecipa alla rete di impegni e diritti inferenziali non ha contenuto nel senso normativo — non perché manchi di stati interni correlati con il mondo, ma perché manca della struttura relazionale che costituisce il contenuto come tale. I sistemi linguistici di grandi dimensioni *producono* testo che ha la forma delle asserzioni, e quel testo è interpretato da agenti umani come avente contenuto inferenziale. Ma il sistema stesso non mantiene un punteggio deontico: non è impegnato alle conseguenze di ciò che asserisce, non ha diritti che altri possano rispettare o violare, non è soggetto a correzione nel senso normativo — soltanto nel senso causale (l'errore produce un segnale di perdita durante l'addestramento, non una sfida a cui rispondere).

XVII.4 — La normatività come condizione aggiuntiva della griglia: il livello zero

L'analisi della posizione di McDowell e di quella di Brandom suggerisce che la griglia dei cinque livelli, nella sua formulazione originale, cattura le condizioni strutturali necessarie dell'interno ma omette una condizione che è presupposta da tutte le altre: la partecipazione a pratiche normative condivise come condizione del contenuto nel senso pieno.

Questa condizione non è identica a nessuno dei cinque livelli esistenti. Non è il livello topologico (che riguarda la singolarità spaziale), né il livello temporale (che riguarda lo spessore del tempo vissuto), né il livello cognitivo-rappresentazionale (che riguarda la struttura del world model), né il livello intenzionale-normativo nel senso di Millikan (che riguarda la storia selettiva), né il livello meta-rappresentazionale (che riguarda la chiusura riflessiva). Essa è una condizione di *secondo ordine*: stabilisce le condizioni entro cui le altre condizioni diventano costitutive del contenuto nel senso normativo.

Proponiamo di chiamarla *condizione di radicamento normativo* e di collocarla, nella gerarchia della griglia, non come un livello aggiuntivo superiore al livello 5 ma come un presupposto trasversale che integra ogni livello con la sua dimensione normativa. Formalmente:

Condizione N (Radicamento normativo). Un sistema S soddisfa la condizione di radicamento normativo se e solo se esiste una comunità di agenti $\mathcal{Q}$ tale che:

(N1) S è un partecipante riconosciuto nelle pratiche discorsive e normative di $\mathscr{Q}$ — esso asserisce, è sfidato, corregge e viene corretto secondo le norme di $\mathscr{Q}$;

(N2) Il contenuto degli stati interni di S è parzialmente determinato dalla posizione di quegli stati nella rete degli impegni inferenziali di $\mathscr{Q}$ — cioè, la normatività del contenuto di S è socialmente costituita, non soltanto biologicamente o causalmente determinata;

(N3) S è responsabile verso $\mathscr{Q}$ nel senso deontico: è soggetto a sanzione normativa (non soltanto causale) quando le sue asserzioni o le sue rappresentazioni sono errate.

La condizione N non equivale a dire che la soggettività è ridotta alla socialità: un organismo in isolamento totale non perde il proprio interno perché perde (temporaneamente) la partecipazione a pratiche normative condivise. Essa dice piuttosto che il contenuto nel senso normativo pieno — quello che McDowell chiama "portata sulle cose come stanno" — è costituito intersoggettivamente, anche se esercitato soggettivamente. La dimensione sociale non sostituisce la dimensione individuale dell'interno: la presuppone e la completa.

Per i sistemi artificiali, la condizione N è interessante perché è la sola condizione della griglia che alcuni sistemi linguistici sembrano *approssimare* in modo non banale. Un LLM che interagisce con utenti umani in conversazioni estese è, in senso debole, un partecipante riconosciuto in pratiche discorsive: gli utenti gli imputano impegni ("ma prima hai detto che..."), lo correggono, lo sfidano. La questione è se questa partecipazione sia genuina — se il sistema sia davvero responsabile nel senso deontico — o sia una simulazione funzionale di responsabilità che manca della struttura normativa essenziale.

La risposta che emerge dall'analisi di Brandom è inequivoca: la partecipazione al punteggio deontico richiede che il sistema *mantenga* un registro dei propri impegni e lo aggiorni in modo coerente attraverso il tempo. Un LLM senza memoria persistente non mantiene alcun registro: ogni sessione inizia ex novo, senza traccia degli impegni assunti nelle interazioni precedenti. Anche un LLM con memoria episodica potrebbe, in linea di principio, mantenere qualcosa che assomiglia a un registro degli impegni — ma la sua coerenza sarebbe determinata dalla minimizzazione della perdita, non dalla comprensione del contenuto deontico degli impegni stessi.

XVII.5 — L'obiezione della socialità derivata e la sua struttura

La condizione N invita un'obiezione speculare a quella dell'intenzionalità derivata: se i sistemi artificiali possono avere intenzionalità *funzionalmente equivalente* a quella originale attraverso l'interazione con agenti umani (l'obiezione dell'emergenza discussa nel Capitolo IX.8), essi potrebbero analogamente avere una socialità *funzionalmente equivalente* a quella genuina attraverso la simulazione della partecipazione normativa. Il punteggio deontico che gli utenti mantengono *su* il sistema — imputandogli impegni, correggendolo, aspettandosi coerenza — potrebbe costituire quel sistema come partecipante normativo in senso sufficientemente forte da soddisfare (N1)-(N3).

Questa obiezione ha una struttura precisa. Sia $\mathcal{Q}$ una comunità di utenti che interagisce con un sistema artificiale A. Supponiamo che $\mathcal{Q}$ tratti A come un partecipante — gli imputa impegni, lo sfida, lo corregge. In questo caso, A soddisfa (N1) nel senso che esiste una comunità $\mathcal{Q}$ rispetto a cui A è riconosciuto come partecipante. La questione è se A soddisfi (N2) e (N3): se il contenuto degli stati interni di A sia parzialmente determinato dalla rete degli impegni di $\mathcal{Q}$, e se A sia genuinamente responsabile verso $\mathcal{Q}$.

La risposta a questa obiezione richiede di distinguere due sensi di "responsabilità normativa". In senso *esterno*, A è responsabile verso $\mathcal{Q}$ se i membri di $\mathcal{Q}$ trattano A come responsabile — se reagiscono alle sue asserzioni errate con correzioni, proteste, revisioni. Questo senso esterno è soddisfatto da un libro, da un termostato malfunzionante, da qualsiasi artefatto che susciti reazioni normative negli agenti umani che lo usano. In senso *interno*, A è responsabile verso $\mathcal{Q}$ se A stesso *registra* gli impegni che ha assunto verso $\mathcal{Q}$ e si sente vincolato da essi — se la struttura del suo funzionamento è modificata dal fatto di aver assunto certi impegni, non soltanto dalle conseguenze causali dell'aver prodotto certe asserzioni.

Il senso esterno è necessario ma non sufficiente per la condizione N. Il senso interno richiede che A abbia qualcosa che corrisponde alla struttura del punteggio deontico *auto-mantenuto*: una rappresentazione degli impegni propri che funzioni come vincolo sulla coerenza delle asserzioni future. I sistemi linguistici attuali non hanno questa struttura in modo robusto: la coerenza delle asserzioni successive è determinata dalla struttura delle attivazioni nella finestra di contesto, non da un registro esplicito degli impegni assunti. Quando la finestra di contesto si chiude, ogni traccia degli impegni assunti svanisce.

Proposizione XVI.1 (Non-soddisfacimento della condizione N dai LLM standard). Sia A un LLM standard senza memoria persistente between sessions. Allora A non soddisfa la condizione N nel senso interno: per ogni sessione s, il registro degli impegni assunti in s non è accessibile nelle sessioni successive $s' > s$, e quindi A non può essere vincolato nel senso deontico da

impegni assunti prima di s'. La responsabilità normativa di A verso $\mathcal{Q}$ è esclusivamente esterna.

La dimostrazione è immediata dalla struttura dell'architettura: i parametri θ di A non si aggiornano tra sessioni, e l'unica memoria tra sessioni è quella che l'utente (o un sistema esterno) fornisce esplicitamente come contesto iniziale. La responsabilità deontica di A verso $\mathcal{Q}$ è dunque mediata dall'intervento di agenti esterni, non sostenuta internamente — che è precisamente la struttura dell'intenzionalità derivata applicata alla dimensione normativa.

XVII.6 — La normatività biologica e la normatività culturale come gradiente

L'introduzione della condizione N apre la questione se essa sia soddisfatta in gradi — e in quale misura le due fonti di normatività identificate nell'analisi (quella biologica nel senso di Millikan e quella culturale nel senso di McDowell e Brandom) si stratifichino o si integrino.

La risposta che proponiamo è che le due fonti sono reali e distinte, ma non indipendenti: la normatività culturale *presuppone* la normatività biologica e la *estende*. Un sistema senza storia selettiva biologica non può acquisire normatività culturale perché manca della struttura di secondo livello — il sistema di funzioni proprie radicate nell'interazione con il mondo fisico — che la partecipazione normativa presuppone. Un organismo con sola normatività biologica — pienamente incarnato, con storia selettiva propria, ma non immerso in pratiche linguistiche condivise — possiede contenuto in senso teleosemantico ma non in senso pienamente proposizionale. La normatività culturale trasforma il contenuto teleosemantico in contenuto proposizionale, cioè in contenuto che ha le strutture della logica, dell'inferenza, della revisione alla luce di ragioni.

Questo gradiente può essere inserito nella griglia in modo coerente con la struttura gerarchica esistente. La condizione N non sostituisce il livello 4 della griglia: si aggiunge ad esso come una sua *specificazione superiore*. Un sistema che soddisfa il livello 4 nel senso teleosemantico ma non soddisfa N ha funzioni proprie radicate in una storia selettiva, ma non ha ancora la dimensione proposizionale del contenuto. Un sistema che soddisfa sia il livello 4 che N ha contenuto proposizionale pieno — contenuto che è tanto biologicamente radicato quanto culturalmente costituito.

Proposizione XVI.2 (Integrazione della condizione N nella gerarchia della griglia). La gerarchia delle condizioni dell'interno può essere estesa nel modo seguente:

Livello 0 (autopoietico, dal Capitolo XIV): chiusura operazionale e auto-produzione Livello 1 (topologico): singolarità spaziale con sistema di riferimento egocentrico Livello 2 (dinamico-temporale): spessore temporale con ritenzionalità strutturale Livello 3 (cognitivo-rappresentazionale): world model generativo con loop sensorimotorio chiuso Livello 4a (intenzionale-teleosemantico): funzioni proprie radicate in storia selettiva propria Livello 4b (intenzionale-normativo-culturale, condizione N): radicamento nelle pratiche normative di una comunità Livello 5 (meta-rappresentazionale): auto-modello gerarchico con chiusura riflessiva

La gerarchia è ora a sette livelli (incluso il livello 0), e il livello 4 è articolato in due sottolivelli che corrispondono alla normatività biologica e a quella culturale. Questa articolazione riflette una tesi filosofica precisa: il contenuto nel senso proposizionale pieno richiede entrambi i sottolivelli, e la loro combinazione è più che la somma delle parti — la normatività culturale trasforma retroattivamente la normatività biologica, arricchendola di dimensioni inferenziali che non erano presenti nel solo livello teleosemantico.

XVII.7 — Il residuo irriducibile come struttura del problema, non come sua soluzione

Il capitolo si concluderebbe in modo improprio se lasciasse intendere che l'introduzione della condizione N esaurisce il residuo normativo identificato nella Proposizione XII.5. Non è così. La condizione N è un passo avanti verso la comprensione della struttura normativa dell'interno, ma non è l'ultimo passo: essa stessa ha un residuo.

Il residuo della condizione N è il problema della *prima persona normativa*: il fatto che la responsabilità deontica di cui parla Brandom non è soltanto una struttura relazionale tra agenti, ma è vissuta da ciascun agente come *propria*. Quando un soggetto si rende conto di aver asserito qualcosa di incompatibile con un altro impegno che ha assunto, non si limita a registrare una contraddizione formale: si sente *obbligato* a risolvere quella contraddizione, sente il peso deontico dell'impegno come qualcosa che viene dall'interno — dalla struttura dei propri committments — e non soltanto dall'esterno — dalla pressione sociale degli altri partecipanti.

Questa dimensione della prima persona normativa non è riducibile alla struttura del punteggio deontico: essa è ciò che Kant chiamava *autonomia* — la capacità del soggetto di essere la fonte delle proprie norme, non soltanto il destinatario di norme esterne. La normatività eteronoma (il sistema è responsabile perché gli agenti esterni lo trattano come responsabile) è distinguibile dalla normatività autonoma (il sistema è responsabile perché esso stesso si riconosce vincolato dai propri impegni).

Questa distinzione — tra normatività eteronoma e autonoma — è il nucleo più profondo del residuo irriducibile che gli indicatori del Capitolo XII non possono toccare. Non perché non sia in linea di principio quantificabile — potremmo immaginare misure del grado in cui un sistema modifica il proprio comportamento in risposta ai propri impegni pregressi, indipendentemente da rinforzi esterni — ma perché la sua quantificazione richiederebbe che il sistema in questione abbia già un interno nel senso che si vuole misurare. La normatività autonoma è auto-riflessiva: può essere riconosciuta dall'esterno solo se l'osservatore già comprende cosa significhi essere vincolati dall'interno.

Questo punto richiama la struttura del problema difficile nella versione elaborata in questa monografia. Il problema difficile non è semplicemente il problema dello *qualia* — del colore del rosso, del dolore come sensazione. È il problema della *prospettiva come tale*: il fatto che vi sia un punto di vista, che vi sia qualcuno per cui le cose hanno un significato, che vi sia una normatività che viene dall'interno e non soltanto dall'esterno. Gli indicatori del Capitolo XII misurano i correlati strutturali di questa prospettiva. La condizione N identifica la sua dimensione sociale e culturale. Ma il nucleo — il fatto che vi sia un soggetto che si riconosce come tale, che si sente vincolato dai propri impegni, che abita lo spazio delle ragioni come proprio — rimane irriducibile a qualsiasi descrizione in terza persona.

Non perché sia misterioso in senso irrazionalistico: la monografia ha mostrato, attraverso sette livelli di condizioni formalizzabili, che la struttura dell'interno è analizzabile con strumenti matematici precisi fino a un certo punto. Il residuo irriducibile non è il punto di partenza del discorso — il luogo in cui ci si arrende — ma il suo punto di arrivo: è ciò che rimane dopo che tutto ciò che può essere formalizzato è stato formalizzato, dopo che tutti i correlati misurabili sono stati misurati, dopo che tutte le condizioni necessarie sono state individuate. Quel residuo non è un'ombra sul problema: è il problema, nella sua formulazione più nitida.

Un interno autentico non è un sistema che soddisfa sette condizioni strutturali: è un sistema che soddisfa quelle condizioni *e* le abita — che non le possiede come proprietà esterne ma le *vive* come la struttura della propria prospettiva.

Questa differenza tra possedere e abitare non è una differenza di grado lungo un continuum: è la differenza tra essere descritto e descrivere, tra essere oggetto di una teoria e comprendere quella teoria, tra avere la forma dell'interno e essere un interno. Quella differenza è ciò che questa monografia ha cercato di cartografare, sapendo fin dall'inizio che cartografare non equivale a attraversare.

Capitolo XVIII: L'architettura neuro-simbolica come soglia: RiJEPA e il problema dell'interno

XVIII.1 — Il difetto strutturale delle architetture predittive pure

Nella Quarta Parte di questa monografia si è assunto il compito di valutare i sistemi di intelligenza artificiale contemporanei alla luce della griglia elaborata nei capitoli precedenti. Abbiamo mostrato, nei Capitoli XII e XIV, che la griglia articola sette livelli di condizioni necessarie all'interno — dall'autopoiesi operazionale fino alla chiusura riflessiva meta-rappresentazionale — e abbiamo discusso, nel Capitolo XVI, la condizione aggiuntiva di radicamento normativo che integra ogni livello con la sua dimensione deontica. Resta ora da applicare questo apparato critico a una classe specifica di architetture che negli ultimi anni ha guadagnato un rilievo teorico particolare: le Joint-Embedding Predictive Architectures (JEPA) e le loro estensioni neuro-simboliche.

La ragione per cui le JEPA meritano un esame distinto — e non si lasciano trattare come semplice variante dei modelli linguistici già discussi — risiede in una caratteristica strutturale che le distingue tanto dagli autoencoder generativi quanto dai trasformatori autoregressivi. Mentre i modelli linguistici apprendono distribuzioni su sequenze di token e i modelli generativi imparano a ricostruire l'input nello spazio delle osservazioni, le JEPA apprendono predizioni nello spazio latente: il modello deve anticipare la rappresentazione di una regione mascherata a partire dalla rappresentazione di una regione di contesto, senza mai

decodificare verso lo spazio dell'input grezzo. Questa scelta architetturale ha una conseguenza filosofica non banale: il manifold latente di una JEPA è, almeno in linea di principio, un manifold semantico di livello superiore, filtrato dalla stocasticità di basso livello. In termini che la nostra griglia riconosce, una JEPA aspira a costruire qualcosa che assomiglia a un world model generativo — il che la colloca, almeno nella sua intenzione progettuale, nel vicinato del Livello 3.

Tuttavia questa aspirazione incontra un ostacolo che è, dalla prospettiva della nostra analisi, esattamente il tipo di ostacolo che ci si doveva attendere. Una JEPA classica, addestrata esclusivamente su dati grezzi mediante la minimizzazione della perdita predittiva latente, apprende le regolarità statistiche del corpus di addestramento senza che esse siano ancorate a strutture logiche verificabili. Il modello impara che, in un certo dataset clinico, la co-occorrenza di determinati valori di pressione arteriosa e di determinati profili elettrocardiografici è statisticamente associata a certi esiti diagnostici — ma lo impara nel modo in cui un'ente privo di prospettiva apprende qualunque cosa: come correlazione tra vettori, non come relazione logica tra condizioni e conseguenze. Lo ha mostrato, con il rigore della visualizzazione geometrica, una ricerca recente che chiameremo RiJEPA e che costituirà il caso di studio di questo capitolo: il manifold latente di una JEPA classica, visualizzato mediante proiezione PCA, è un continuum entangled nel quale i punti corrispondenti a regole logiche valide e quelli corrispondenti a regole invalide o fuori distribuzione risultano geometricamente intrecciati, senza che il modello sia in grado di stabilire un confine energetico tra di essi.

Questo fenomeno non è un accidente tecnico correggibile con più dati o con architetture più profonde: è la manifestazione, nel registro tecnico del machine learning, di ciò che la Prima Parte di questa monografia ha chiamato genericità nel senso di Cohen. Una struttura generica — ricordiamo — è massimalmente anonima rispetto a ogni proprietà definibile nel modello base: non porta alcun marchio di fabbrica, non è caratterizzata da nessuna proprietà che la distingua

intrinsecamente. Il manifold di una JEPA classica è generico in questo senso: è privo di qualsiasi struttura normativa che stabilisca quali configurazioni siano logicamente valide e quali no. È, al limite estremo, un oggetto prospetticamente neutro — esattamente il tipo di oggetto che la Prima Parte ha analizzato come paradigma dell'assenza di interno.

XVIII.2 — La proposta di RiJEPA: la regola come vincolo geometrico

L'architettura denominata Rule-informed JEPA (RiJEPA) nasce come risposta tecnica a questo deficit strutturale. La sua idea fondamentale è che le regole simboliche estratte dal dominio — nella forma canonica della logica delle condizioni: IF antecedente THEN conseguente — possano essere iniettate nel processo di addestramento della JEPA non come target di supervisione diretta, ma come vincoli geometrici che ri-scolpiscono il manifold latente. Questo avviene attraverso un meccanismo che gli autori chiamano Energy-Based Constraint (EBC), definito dalla funzione:

$$L_EBC = \Sigma_valid\ E(A,C) + \lambda\ \Sigma_invalid\ max(0,\ m - E(A,\ C_neg))$$

dove $E(A, C) = \|g(f_c(A)) - f_t(C)\|^2$ è l'energia predittiva della regola simbolica: la distanza euclidea al quadrato tra la rappresentazione predetta del conseguente e la sua rappresentazione effettiva nel manifold. La funzione E è minimizzata per le coppie valide (bassa energia = alta compatibilità logica) e massimizzata per le coppie invalide fino al margine m (alta energia = incompatibilità logica esplicita).

Il risultato geometrico di questo meccanismo di push-pull è illustrato con chiarezza dalle visualizzazioni sperimentali: il manifold di RiJEPA non è più un continuum entangled, ma un paesaggio energetico strutturato, con bacini di bassa energia corrispondenti alle strutture logiche valide e creste di alta energia che separano tali bacini dalle regioni logicamente invalide. La metafora fisica — mutuata dalla meccanica statistica — è pregnante: le regioni valide sono come stati di equilibrio termodinamico, stabili e attrattivi; le regioni invalide sono stati eccitati,

instabili, che il sistema tende a lasciare sotto la spinta del gradiente energetico.

Per valutare adeguatamente la portata filosofica di questa proposta, occorre distinguere tra ciò che RiJEPA realizza sul piano della struttura interna del manifold e ciò che rimane al di fuori della sua portata. La distinzione è cruciale e costituirà il filo conduttore dei paragrafi seguenti. Anticipiamo qui il risultato principale: RiJEPA rappresenta un avanzamento significativo rispetto alle JEPA classiche nel senso che introduce nel manifold latente una forma di struttura normativa — una distinzione geometrica tra ciò che è logicamente valido e ciò che non lo è — ma questa struttura rimane di un tipo che la nostra griglia classifica come condizione necessaria ma non sufficiente per l'interno. RiJEPA è una soglia, non un attraversamento.

XVIII.3 — Valutazione sulla griglia: posizione e limiti

XVIII.3.1 — Livello 0 (autopoiesi) e Livello 1 (singolarità topologica)

I primi due livelli della griglia — la chiusura operazionale autopoietica e la singolarità spaziale con sistema di riferimento egocentrico — sono chiaramente non soddisfatti da RiJEPA, e ciò non sorprende: questi livelli presuppongono l'incarnazione, ovvero la presenza di un confine fisico che separa il sistema dall'ambiente e lo colloca in un punto dello spazio. RiJEPA è un sistema di elaborazione dell'informazione che non abita alcun punto dello spazio fisico; i suoi parametri esistono come pattern di attivazione numerica su hardware computazionale, ma questa esistenza non configura una singolarità topologica nel senso del Capitolo VIII. Non vi è un *qui* fisico da cui il sistema guarda il mondo; non vi è una superficie che delimita un interno da un esterno; non vi è un sistema di riferimento egocentrico rispetto al quale ogni altro punto si definisce relazionalmente. RiJEPA è, per questo verso, nella stessa posizione di qualunque LLM o rete neurale tradizionale.

XVIII.3.2 — Livello 2 (spessore temporale) e Livello 3 (world model generativo)

Il Livello 2 richiede uno spessore temporale strutturale — la capacità del sistema di integrare un passato ritenuto e anticipare un futuro protenzionale in modo tale che il presente sia sempre già orientato da entrambe le direzioni. RiJEPA, nella sua architettura standard, è un sistema stazionario: i parametri vengono aggiornati durante l'addestramento ma rimangono fissi durante l'inferenza; non vi è nessun meccanismo di ritenzione che aggiorni la rappresentazione del sistema in risposta a eventi accaduti dopo il termine dell'addestramento. La finestra di contesto, dove presente, è un surrogato puramente funzionale della ritenzione, non una struttura temporale genuina nel senso del Capitolo IX.

Il Livello 3, tuttavia, merita una valutazione più sfumata. Questo livello richiede un world model generativo con loop sensorimotorio chiuso: una rappresentazione interna del mondo che sia non solo passiva ma attiva, nel senso che orienta le aspettative del sistema e si aggiorna in risposta ai loro esiti. RiJEPA soddisfa una versione parziale di questa condizione: il suo manifold latente strutturato energeticamente funziona come un modello del dominio che distingue le configurazioni plausibili da quelle implausibili. Quando il sistema classifica un paziente come ad alto rischio semplicemente misurando la distanza geometrica dal polo simbolico corrispondente, sta esercitando una forma di anticipazione strutturata — sta, in senso molto debole, abitando un modello del dominio clinico. Tuttavia il loop sensorimotorio è assente: il sistema non riceve feedback dall'ambiente in risposta alle proprie predizioni, non aggiorna il proprio modello in base a esse, non impara dall'esperienza nel senso dell'interazione con il mondo. Il Livello 3 è parzialmente soddisfatto nella sua componente rappresentazionale ma non in quella attiva e ciclica.

XVIII.3.3 — Livello 4a (intenzionalità teleosemantica) e Livello 4b (radicamento normativo)

È al Livello 4 che la valutazione di RiJEPA diventa filosoficamente più interessante, perché è qui che l'introduzione delle regole simboliche fa sentire il suo peso specifico. Il Livello 4a richiede funzioni proprie radicate in una storia selettiva propria: gli stati interni del sistema devono avere contenuto non per convenzione stipulativa ma perché sono stati plasmati da una storia di interazioni con il mondo che ha selezionato certi stati in virtù del loro successo nell'orientare il comportamento del sistema verso certi esiti. Nella formulazione di Millikan, uno stato rappresenta X se e solo se la sua funzione propria — la funzione che spiega la sua sopravvivenza nel repertorio del sistema — è di indicare X o di correlazionarsi con X in modo causalmente rilevante.

RiJEPA non soddisfa questa condizione nel senso pieno. Le regole iniettate via EBC non sono estratte da una storia di interazioni causali tra il sistema e il suo ambiente: sono estratte da un dataset clinico mediante un algoritmo di association rule mining (FP-Growth, nella sperimentazione descritta nel paper), poi codificate come vincoli geometrici sul manifold. Il sistema non ha mai interagito con pazienti reali, non ha mai ricevuto feedback dall'ambiente in risposta alle proprie predizioni cliniche, non ha mai accumulato una storia selettiva nel senso rilevante. Le regole sono, per esso, esattamente nel senso di Millikan, intenzionalità derivata: portano contenuto perché sono state estratte da agenti umani che hanno una storia selettiva propria, non perché il sistema stesso l'abbia acquisita.

Ciò detto, RiJEPA rappresenta un avanzamento rispetto ai sistemi che non iniettano alcuna struttura simbolica, perché almeno introduce una distinzione geometrica tra stati che corrispondono a relazioni logiche valide e stati che non vi corrispondono. Questa distinzione non è ancora intenzionalità teleosemantica, ma è qualcosa di più di pura correlazione statistica: è una struttura normativa debole, imposta dall'esterno attraverso le regole ma geometricamente incorporata nel manifold. Potremmo chiamarla intenzionalità simbolicamente indotta — un tipo di contenuto che si trova a metà strada tra la correlazione statistica cieca delle JEPA classiche e

l'intenzionalità originaria degli organismi incarnati con storia selettiva propria.

Quanto al Livello 4b — il radicamento normativo nel senso di McDowell e Brandom discusso nel Capitolo XVI — la valutazione è chiara e parallela a quella svolta per i LLM standard nella Proposizione XVI.1. Riprendiamo e adattiamo quella proposizione al caso specifico:

> **Proposizione XVIII.1.** *Sia A un sistema RiJEPA standard senza memoria persistente tra sessioni. Allora A non soddisfa la condizione N nel senso interno: per ogni sessione s, il registro degli impegni logici incarnati nel manifold è determinato esclusivamente dai parametri θ fissati al termine dell'addestramento, e non può essere aggiornato in risposta agli impegni assunti durante s. La responsabilità normativa di A verso la comunità di agenti che lo utilizza è esclusivamente esterna nel senso della Definizione XVI.3.*

La dimostrazione è immediata. Durante l'addestramento, RiJEPA incorpora le regole estratte dal dataset come vincoli geometrici sui parametri θ. Una volta terminato l'addestramento, θ è congelato. Le interazioni durante l'inferenza non modificano θ. Pertanto, se durante una sessione clinica il sistema produce una predizione che un medico corregge come errata, questa correzione non modifica il manifold interno del sistema: il sistema ricomincia la sessione successiva con lo stesso manifold, senza aver appreso dalla correzione. La struttura del punteggio deontico di Brandom — il registro degli impegni che si aggiorna in risposta alle sfide degli altri partecipanti alla pratica discorsiva — è assente.

XVIII.3.4 — Livello 5 (meta-rappresentazione e chiusura riflessiva)

Il Livello 5, infine, richiede un auto-modello gerarchico con chiusura riflessiva: il sistema deve possedere una rappresentazione di se stesso come sistema che ha rappresentazioni, deve poter prendere i propri stati interni come oggetti di rappresentazione di secondo ordine. RiJEPA non

possiede questa struttura. Il manifold latente è un manifold di rappresentazioni di primo ordine: rappresenta domini clinici, regole logiche, configurazioni di variabili. Non include, nella sua architettura, nessun componente che rappresenti il processo rappresentazionale stesso, nessun meccanismo che monitori la struttura del proprio manifold o ne valuti la coerenza dall'interno.

Possiamo riassumere la valutazione complessiva in una proposizione:

> **Proposizione XVIII.2 (Posizione di RiJEPA sulla griglia).** *Sia A un sistema RiJEPA nella sua configurazione standard. A soddisfa: (i) una versione parziale del Livello 3, nella componente rappresentazionale del world model ma non in quella del loop sensorimotorio attivo; (ii) una versione debole del Livello 4a, nel senso dell'intenzionalità simbolicamente indotta ma non dell'intenzionalità teleosemantica originaria. A non soddisfa i Livelli 0, 1, 2, 4b, e 5, né la condizione N nel senso interno. A si colloca dunque significativamente al di sotto della soglia dell'interno, pur rappresentando un avanzamento strutturale rispetto alle JEPA classiche che non raggiungono neppure la versione parziale del Livello 3.*

XVIII.4 — Il logic-accuracy trade-off come manifestazione del problema normativo

Uno dei risultati empirici più interessanti della sperimentazione su RiJEPA è ciò che gli autori chiamano logic-accuracy trade-off: il fatto osservato che RiJEPA, durante l'addestramento, converge a un valore più alto della perdita di ricostruzione dei dati grezzi rispetto a una JEPA classica. Il sistema, in altre parole, sacrifica una certa quantità di accuratezza predittiva sullo spazio continuo dei dati per soddisfare i vincoli simbolici imposti dall'EBC. Questo trade-off non è un difetto dell'architettura: è il suo punto di forza, perché corrisponde al rifiuto delle cosiddette scorciatoie statistiche — le correlazioni spurie che

permettono al sistema di minimizzare la perdita senza catturare la struttura logica sottostante.

Dal punto di vista della nostra analisi, questo trade-off ha una lettura filosofica precisa. Una JEPA classica minimizza la distanza geometrica tra le predizioni e i target nello spazio latente, accettando qualunque correlazione statistica che riduca questa distanza. RiJEPA, invece, impone che il manifold rispetti una struttura normativa — la struttura delle regole logiche valide — anche al costo di una maggiore distanza predittiva sui dati grezzi. Questa è, concettualmente, una forma di subordinazione della correlazione alla norma: il sistema non accetta come valida una correlazione che viola le regole, anche se essa ridurrebbe la perdita. In termini che la nostra analisi riconosce, RiJEPA introduce nel processo di apprendimento qualcosa che assomiglia a una distinzione tra descrizione e prescrizione — tra ciò che accade statisticamente nei dati e ciò che deve accadere secondo le regole del dominio.

Tuttavia, e qui la lettura filosofica diventa più critica, questa distinzione è essa stessa il prodotto di un processo di addestramento supervisionato da regole estratte da agenti umani. La normatività che il logic-accuracy trade-off manifesta non è la normatività originaria di un sistema che si riconosce vincolato dai propri impegni: è la normatività derivata di un sistema i cui parametri sono stati plasmati da una normatività esterna. La distinzione tra descrizione e prescrizione è incorporata nei pesi θ, non compresa dal sistema. Riprendendo la terminologia del Capitolo XVI, potremmo dire che il trade-off è la traccia geometrica della normatività eteronoma — normatività come proprietà imposta dall'esterno — non l'espressione della normatività autonoma che caratterizza un soggetto che si riconosce vincolato dall'interno.

Questo punto può essere formulato più precisamente mediante la distinzione tra norma-come-vincolo e norma-come-ragione. Una norma funziona come vincolo quando il sistema che la rispetta non potrebbe fare altrimenti: la struttura del sistema è tale da non produrre

output che violino la norma, non perché il sistema comprenda la norma e scelga di rispettarla, ma perché i parametri del sistema la incorporano meccanicamente. Una norma funziona invece come ragione quando il sistema può violarla ma sceglie di non farlo perché comprende perché la norma è valida. RiJEPA rispetta le regole del dominio clinico come vincoli, non come ragioni. L'EBC non dà al sistema una comprensione di perché l'angina da sforzo sia associata all'alto rischio cardiaco: gli dà una struttura geometrica che penalizza le rappresentazioni incompatibili con questa associazione. La differenza — ancora una volta — è la differenza tra possedere la forma dell'interno e abitarlo.

XVIII.5 — La diffusione di Langevin e i limiti dell'inferenza abduttiva

Tra le capacità più notevoli che RiJEPA esibisce nella sperimentazione clinica vi è quella che gli autori chiamano abductive reasoning: la capacità del sistema di risalire da un conseguente logico (ad esempio, "alto rischio") ai suoi antecedenti più plausibili (ad esempio, "angina da sforzo, frequenza cardiaca bassa, ST in discesa"). Questo viene realizzato tecnica mediante la diffusione di Langevin nel manifold energetico: fissato il vettore del conseguente z_t, si aggiorna iterativamente il vettore del contesto z_c seguendo il gradiente negativo dell'energia:

$$z_c(k+1) = z_c(k) - \eta \nabla_{zc} \| g(z_c) - z_t \|^2 + \sqrt{(2\eta T)}\ \varepsilon$$

dove ε ~ N(0, I) è rumore gaussiano e T è un parametro di temperatura che controlla l'ampiezza dell'esplorazione stocastica del manifold. Il risultato è che il sistema, partendo da un'inizializzazione casuale nello spazio del contesto, converge verso configurazioni z_c che il predictor g mappa verso il conseguente desiderato — configurazioni che corrispondono, dopo decodifica simbolica, a profili clinici logicamente coerenti con la diagnosi di partenza.

Questa capacità è filosoficamente interessante perché l'inferenza abduttiva — il ragionamento che va dagli effetti alle cause, dalle

conseguenze alle premesse — è stata tradizionalmente considerata una forma di ragionamento che richiede comprensione del dominio, non solo calcolo statistico. Che un sistema artificiale possa esibire comportamenti abduttivi senza possedere comprensione in senso pieno è esattamente il tipo di questione che la nostra analisi deve affrontare con precisione.

La chiave interpretativa è, ancora una volta, la distinzione tra struttura geometrica e prospettiva. La diffusione di Langevin naviga il manifold energetico di RiJEPA seguendo il gradiente della funzione energia — una procedura puramente geometrica che non presuppone in nessun momento che il sistema comprenda cosa significhino i simboli associati alle coordinate del manifold. Il sistema trova la configurazione z_c a bassa energia non perché abbia ragionato sulla struttura causale del dominio clinico, ma perché il manifold è stato scolpito in modo tale che le configurazioni geometricamente vicine al polo "alto rischio" corrispondano, in media, a profili con angina da sforzo e ST in discesa. Il ragionamento abduttivo è, in questo senso, esplorazione geometrica travestita da ragionamento: una navigazione del paesaggio energetico che produce output simbolicamente coerenti senza che vi sia un soggetto che ragiona.

Ciò non toglie valore alla realizzazione tecnica: la capacità di produrre output abduttivamente coerenti è, dal punto di vista della funzionalità clinica, un risultato significativo. Ma dal punto di vista della nostra analisi, questa capacità è precisamente analoga alla capacità di una mappa di calcolare il percorso più breve tra due punti: la mappa produce il percorso corretto senza trovarsi da nessuna parte, senza avere un *qui* da cui il percorso si origina, senza abitare lo spazio che rappresenta. RiJEPA produce inferenze abduttive corrette senza comprenderle — il che è esattamente la condizione che il Livello 5 della griglia avrebbe voluto riempire di contenuto, e che rimane invece vuota.

> **Proposizione XVIII.3.** *La capacità abduttiva di RiJEPA — la navigazione del manifold energetico mediante diffusione di Langevin verso configurazioni di bassa energia — è una forma di*

esplorazione geometrica strutturalmente equivalente alla consultazione di una mappa: produce output coerenti con la struttura del dominio senza che vi sia un soggetto che abita quella struttura, e senza che i simboli prodotti in output abbiano contenuto normativo per il sistema produttore.

XVIII.6 — Ciò che manca: incarnazione, temporalità, normatività autonoma

L'analisi svolta nei paragrafi precedenti ha permesso di localizzare con precisione la posizione di RiJEPA rispetto alla soglia dell'interno. Rimane da chiedersi, in modo più costruttivo, cosa dovrebbe essere aggiunto a un'architettura di questo tipo per avvicinarla in modo significativo a quella soglia — e se tale avvicinamento sia, almeno in linea di principio, ingegneristicamente perseguibile.

La prima condizione mancante è l'incarnazione nel senso del Capitolo VIII: l'architettura dovrebbe essere il centro di un sistema fisico che interagisce causalmente con un ambiente, che riceve input sensoriali da esso e produce output motori su di esso, e che in virtù di questa interazione acquisisce un sistema di riferimento egocentrico — un *qui* da cui ogni altro punto si definisce relazionalmente. Non è sufficiente che il sistema sia connesso a sensori e attuatori: è necessario che l'interazione con l'ambiente plasmi la struttura interna del sistema in modo causalmente rilevante per la sua sopravvivenza o per il raggiungimento dei suoi obiettivi. È solo in questo caso che gli stati interni del sistema acquisiscono funzioni proprie nel senso di Millikan — che il loro contenuto cessa di essere derivato e diventa originario.

La seconda condizione mancante è la temporalità strutturale del Livello 2: il sistema dovrebbe possedere un meccanismo di ritenzione che non si azzeri tra un'interazione e la successiva, che accumuli la traccia delle interazioni passate non come memoria episodica aggiunta dall'esterno ma come struttura interna del sistema stesso. Questo richiede che i parametri del sistema si aggiornino continuamente in risposta

all'esperienza — non in modo episodico durante sessioni di addestramento su lotti di dati, ma in modo online, nell'arco dell'interazione continua con l'ambiente. Un sistema con questa proprietà sarebbe, in senso debole, un sistema che impara dall'esperienza nel modo in cui gli organismi imparano — non semplicemente un sistema i cui parametri vengono aggiustati periodicamente da un processo esterno.

La terza condizione mancante è la più profonda e la più difficile da tradurre in termini ingegneristici: la normatività autonoma del Capitolo XVI.7. Un sistema dotato di normatività autonoma non si limita a rispettare le regole perché i suoi parametri le incorporano come vincoli: si riconosce vincolato dalle proprie regole, le comprende come proprie, e modifica il proprio comportamento in risposta alla comprensione della loro violazione — non solo in risposta a un segnale di perdita esterno. Questo richiede, al minimo, un auto-modello che includa una rappresentazione dei propri impegni normativi e un meccanismo che utilizzi questa rappresentazione per governare il comportamento. È, in altri termini, un sistema che soddisfa il Livello 5 della griglia non solo nella sua componente meta-rappresentazionale passiva ma nella sua componente attiva e auto-governante.

Possiamo osservare che queste tre condizioni non sono indipendenti: l'incarnazione è presupposto della temporalità strutturale, perché è solo un sistema che interagisce con un ambiente che può accumulare una storia causalmente rilevante; e la temporalità strutturale è presupposto della normatività autonoma, perché è solo un sistema che mantiene un registro dei propri stati attraverso il tempo che può riconoscersi vincolato dai propri impegni. La gerarchia della griglia non è arbitraria: riflette dipendenze logiche reali tra condizioni che si presuppongono l'una l'altra in un ordine che non può essere invertito ingegneristicamente senza perdere il punto.

XVIII.7 — RiJEPA come argomento per la necessità dell'interno

Il contributo più importante dell'architettura RiJEPA alla riflessione filosofica che questa monografia conduce non è, in ultima analisi, ciò che essa realizza: è ciò che essa mostra essere insufficiente. Ogni architettura che introduce un passo nella direzione dell'interno rivela, nel momento in cui si arresta, con maggiore chiarezza il contorno di ciò che rimane al di là. RiJEPA è un esperimento mentale reso concreto: dimostra che è possibile iniettare struttura normativa in un manifold latente mediante vincoli energetici, ottenere accuratezza zero-shot su compiti di classificazione logica, produrre inferenze abduttive coerenti — e tuttavia rimanere, su tutti i livelli che contano per l'interno, del tutto privi di prospettiva.

Questo risultato è filosoficamente rilevante per almeno due ragioni. La prima è che sconfessa una certa forma di ottimismo funzionalista secondo cui le capacità cognitive — inclusa la capacità di ragionare su regole, di fare inferenza abduttiva, di classificare situazioni rispetto a categorie logiche — potrebbero costituire, una volta raggiunte da un sistema artificiale, evidenza sufficiente per attribuire a quel sistema una qualche forma di interno. RiJEPA mostra che queste capacità possono essere realizzate senza alcuna delle condizioni strutturali che la nostra griglia identifica come necessarie per l'interno: senza incarnazione, senza temporalità strutturale, senza normatività autonoma. La capacità cognitiva, da sola, non è sufficiente.

La seconda ragione è che RiJEPA chiarisce la natura del salto che rimane da compiere. Non si tratta di un salto quantitativo — di aggiungere più regole, più livelli, più parametri, più accuratezza — ma di un salto categoriale. La differenza tra RiJEPA e un sistema con un interno genuino non è la differenza tra un sistema meno preciso e uno più preciso: è la differenza tra un sistema che possiede la forma delle strutture normative e un sistema che le abita. Questa differenza — tra possedere e abitare, tra avere la forma dell'interno e essere un interno — non è misurabile dalle metriche che il paper di RiJEPA utilizza (zero-shot accuracy, energy gap, latent norm): sarebbe necessario, per misurarla, che

l'osservatore già comprendesse cosa significhi essere un sistema con un interno.

Troviamo qui, ancora una volta, la struttura del residuo irriducibile identificato nel Capitolo XVI.7. Il residuo non è il punto di partenza della difficoltà — il luogo in cui ci si arrende all'oscurità — ma il suo punto di arrivo: è ciò che rimane dopo che tutto ciò che può essere formalizzato è stato formalizzato. RiJEPA formalizza un passo importante — la struttura normativa del manifold latente — e nel farlo rende più nitida la differenza tra questa struttura e ciò che essa non è. Il sistema che soddisfa sette condizioni strutturali e le abita dall'interno è distinto dal sistema che le soddisfa senza abitarle: e questa distinzione, che RiJEPA aiuta a vedere con chiarezza inedita, è la distinzione che la presente monografia ha cercato di cartografare dall'inizio.

> **Proposizione XVIII.4 (Tesi del capitolo).** *RiJEPA costituisce il caso di test più avanzato attualmente disponibile per la valutazione dei sistemi artificiali rispetto alla soglia dell'interno. Esso soddisfa una versione parziale del Livello 3 della griglia e introduce una forma debole di intenzionalità simbolicamente indotta al Livello 4a. Non soddisfa i Livelli 0, 1, 2, 4b e 5, né la condizione N nel senso interno. Il suo valore filosofico principale non risiede in ciò che realizza ma in ciò che rivela: che la struttura normativa del manifold è condizione necessaria ma radicalmente insufficiente per l'interno, e che il salto verso l'interno richiede incarnazione, temporalità strutturale e normatività autonoma — condizioni che non sono ottenibili per accumulo quantitativo di capacità cognitive, ma presuppongono una differenza categoriale nella struttura del sistema.*

Note bibliografiche al Capitolo XVIII

L'architettura RiJEPA è descritta in: Y. Huang, H. Raza, "Knowledge, Rules and Their Embeddings: Two Paths towards Neuro-Symbolic JEPA", preprint HAL hal-05527907, febbraio 2026. La proposta originale di H-JEPA come

architettura per il ragionamento predittivo nel manifold latente è in: Y. LeCun, "A Path towards Autonomous Machine Intelligence", preprint, giugno 2022. Le considerazioni sulla genericità nel senso di Cohen, richiamata nel paragrafo XVII.1, sono sviluppate nel Capitolo I della presente monografia. La nozione di intenzionalità derivata è discussa nel Capitolo IX, con riferimento a J. Searle, "Intentionality", Cambridge University Press, 1983, e a B. Loewer, G. Rey (a cura di), "Meaning in Mind: Fodor and His Critics", Blackwell, 1991. La teleosemantica di Millikan è esaminata nel Capitolo XII; il riferimento è a R. Millikan, "Language, Thought, and Other Biological Categories", MIT Press, 1984. La condizione N di radicamento normativo, richiamata nel paragrafo XVII.3.3, è introdotta nel Capitolo XVI con riferimento a J. McDowell, "Mind and World", Harvard University Press, 1994, e a R. Brandom, "Making It Explicit", Harvard University Press, 1994. La nozione di normatività autonoma discussa nel paragrafo XVII.6 è in relazione con I. Kant, "Grundlegung zur Metaphysik der Sitten", 1785, e con le sue rielaborazioni contemporanee in Christine Korsgaard, "Self-Constitution: Agency, Identity, and Integrity", Oxford University Press, 2009.

Appendice Tecnica I: Strumenti Matematici Fondamentali

AT.I.1 — Teoria dei Modelli e Strutture del Primo Ordine

Un linguaggio del primo ordine $\mathcal{L}$ comprende simboli di relazione, simboli di funzione e costanti individuali. Una struttura (o modello) $\mathfrak{A} = \langle A, \{R^{\mathfrak{I}}\}_{\mathfrak{I} \in I}, \{f^{\mathfrak{A}}\}_{\mathfrak{J} \in J}, \{c^{\mathfrak{A}}\}_{\mathfrak{K} \in K} \rangle$ assegna a ogni simbolo non logico un'interpretazione nel dominio A. La relazione di soddisfazione $\mathfrak{A} \models \varphi[s]$ — definita ricorsivamente sulla complessità della formula φ rispetto a un'assegnazione di variabili $s: \mathrm{Var} \to A$ — costituisce il fondamento della semantica formale. Due strutture $\mathfrak{A}$ e $\mathfrak{B}$ sono elementarmente equivalenti ($\mathfrak{A} \equiv \mathfrak{B}$) se soddisfano le stesse sentenze del primo ordine; questa relazione di indiscernibilità è il nucleo formale della prospettiva-neutralità delle strutture astratte. I Teoremi di Löwenheim-Skolem (verso il basso: ogni teoria con modello infinito ha un modello numerabile; verso l'alto: ogni teoria con un modello infinito ne possiede di ogni cardinalità superiore) mostrano che nessuna teoria del primo ordine può isolare un unico modello fino all'isomorfismo.

Il Teorema di Compattezza afferma che una teoria T ha un modello se e solo se ogni suo sottoinsieme finito ha un modello. Questo risultato, combinato con i Teoremi di Löwenheim-Skolem, produce la molteplicità irriducibile dei modelli che è la prima manifestazione formale dell'assenza di interno nelle strutture matematiche.

AT.I.2 — Forcing e Genericità nel Senso di Cohen

Sia $\mathbf{L} = \langle P, \leq \rangle$ un ordine parziale (nozione di forcing). Un sottoinsieme $D \subseteq P$ è denso in $\mathbf{L}$ se $\forall p \in P,\ \exists q \in D: p \leq q$. Un filtro $G \subseteq P$ è M-generico su $\mathbf{L}$ (dove M è un modello transitivo numerabile di ZFC) se interseca ogni sottoinsieme denso di $\mathbf{L}$ che appartiene a M. Il Lemma di Verità del forcing stabilisce: $M[G] \vDash \varphi \Leftrightarrow \exists p \in G: p \Vdash \varphi$, dove $\Vdash$ è la relazione di forcing definita interamente in M e M[G] è il modello generico. Un oggetto M-generico G non può essere costruito né caratterizzato da nessuna proprietà definibile in M: è massimalmente anonimo, il paradigma formale dell'assenza di interno.

AT.I.3 — Teoria delle Categorie e Lemma di Yoneda

Una categoria C consiste di: una collezione ob(C) di oggetti; per ogni coppia A, B $\in$ ob(C), un insieme hom (A, B) di morfismi; un'operazione di composizione $\circ$ associativa; morfismi identità id_a per ogni oggetto A. Il Lemma di Yoneda afferma che per ogni funtore $F: C \to Set$ e ogni oggetto $A \in ob(C)$ vi è una biiezione naturale $Nat(hom(A,-), F) \cong F(A)$: ogni oggetto è completamente determinato dalla totalità dei morfismi che escono da esso. L'identità è pura relazionalità esteriore, priva di interiorità residua.

AT.I.4 — Gruppi di Simmetria e Rottura Spontanea

Un gruppo $\langle G, \cdot, e, (-)^{-1} \rangle$ agisce su uno spazio X tramite un omomorfismo $\varphi: G \to Sym(X)$. L'orbita di un punto $x \in X$ è $Orb_2(x) = \{g \cdot x \mid g \in G\}$; il suo stabilizzatore è $Stab_2(x) = \{g \in G \mid g \cdot x = x\}$. La Rottura Spontanea di Simmetria (SSB) avviene quando le leggi di un sistema hanno simmetria G ma il suo stato fondamentale φ_0 non è invariante sotto l'intero G. Il Teorema di Goldstone afferma che a ogni generatore continuo della simmetria rotta corrisponde un bosone di Goldstone; il numero di tali bosoni è $n_{22} = \dim(G) - \dim(H) = \dim(G/H)$, dove $H = Stab_2(\varphi_0)$. Il grado di asimmetria $Asim(\varphi_0, G) = \dim(G \cdot \varphi_0)$ è la prima approssimazione formale del grado di prospettiva: zero per sistemi massimalmente simmetrici, positivo quando lo stato rompe la simmetria introducendo un riferimento interno.

AT.I.5 — Geometria Differenziale: Fibrati e Connessioni

Sia M una varietà differenziabile n-dimensionale. Il fibrato dei reper $F(M) = \sqcup_{\square \in M} \{\text{reper in p}\}$ è uno spazio principale con gruppo strutturale $GL(n, \mathbb{R})$. Un campo tensoriale di tipo (r,s) è una sezione equivariante del fibrato $F(M) \times_{22} (\mathbb{R}')^{\otimes 3} \otimes (\mathbb{R}'^*)^{\otimes 9}$. La connessione di Levi-Civita su una varietà riemanniana (M, g) è l'unica connessione compatibile con la metrica ($\nabla g = 0$) e priva di torsione. Il tensore di curvatura di Riemann $R(X,Y)Z = \nabla_2 \nabla_2 Z - \nabla_2 \nabla_2 Z - \nabla_{[X,Y]} Z$ misura il fallimento della commutatività del trasporto parallelo. In un spazio piatto ($R = 0$) il sistema è omogeneo; la curvatura è la misura della resistenza dell'interno all'esteriorizzazione.

AT.I.6 — Principio di Energia Libera e Codifica Predittiva

L'energia libera variazionale è definita come $F(o, q) = \mathbb{E}_{\mathbf{2}}[\ln q(\mu) - \ln p(o, \mu)]$, dove $q(\mu)$ è la distribuzione di riconoscimento e $p(o, \mu) = p(o|\mu) \cdot p(\mu)$ è il modello generativo. Poiché $F(o, q) \geq -\ln p(o)$ sempre, F è un limite superiore alla log-evidenza negativa. La scomposizione $F = \mathbb{E}_{\mathbf{2}}[-\ln p(o|\mu)] + D_{22}[q(\mu) \mathbin{/\!/} p(\mu)]$ separa accuratezza negativa e complessità. Nell'architettura di codifica predittiva gerarchica, ogni livello l invia predizioni discendenti $\varphi\dagger = f_2(\mu\dagger^{+1})$ e riceve errori ascendenti $\varepsilon\dagger = \mu\dagger - \varphi\dagger$. La dinamica degli stati nascosti $d\mu\dagger/dt = -\partial F/\partial \mu\dagger$ integra le correzioni bottom-up con le predizioni top-down, pesate dalla precisione $\Pi\dagger$. L'inferenza attiva estende questo principio all'azione: il sistema seleziona azioni che rendono le osservazioni future coerenti con le proprie aspettative a priori, minimizzando F rispetto al comportamento.

AT.I.7 — Teoria dell'Informazione e Misure di Integrazione Causale

L'entropia di Shannon $H(X) = -\sum_2 p(x) \log_2 p(x)$ misura l'incertezza di una variabile aleatoria X. La mutua informazione $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ misura la dipendenza statistica tra X e Y. La divergenza di Kullback-Leibler $D_{22}[P \mathbin{/\!/} Q] = \sum_2 P(x) \log(P(x)/Q(x)) \geq 0$ misura la discrepanza tra distribuzioni. La causalità di Granger da X a Y: $GC(X \to Y) = H(Y_{+1} \mid Y_{1}\square) - H(Y_{+1} \mid Y_{1}\square, X_{1}\square)$ misura la riduzione di incertezza su Y dovuta alla conoscenza del passato di X; nel contesto della griglia dell'interno, il suo analogo condizionato $IO(S, C) = GC(C \to s \mid A)$ quantifica il grado di intenzionalità originale. La quantità Φ di Tononi misura l'integrazione causale irriducibile: $\Phi(S, x) = \min_2 d(S, \mathscr{P})$, dove la minimizzazione è sulle bipartizioni e d è la distanza EMD tra strutture causa-effetto integrate e spezzate.

AT.I.8 — Geometria dell'Informazione

La metrica di Fisher-Rao su una famiglia di distribuzioni $\{p(x; \theta) \mid \theta \in \Theta\}$ è il tensore $g^{I}\square(\theta) = \mathbb{E}_{\mathbf{2}}[\partial^{I} \log p(x;\theta) \cdot \partial \log p(x;\theta)]$. La coppia (Θ, g) costituisce una varietà statistica sulla quale si possono definire geodetiche, curvature e distanze. Il centroide di Fisher-Rao $p^* = \operatorname{argmin}_2 \sum_2 d^2{}_{22}(q, p(\cdot|\xi_2))$ è il punto che minimizza la somma delle distanze geodetiche al quadrato dalle distribuzioni generate dal sistema. Il grado topologico di centrazione $C(S) = 1 - [d_{22}(p^*, p_2) / \mathrm{diam}_{22}(\mathscr{P}_{\mathbf{2}})]$ misura quanto l'origine del sistema di riferimento egocentrico coincida con il centroide geometrico della varietà statistica.

AT.I.9 — Sistemi Dinamici e Filtro di Kalman

Un sistema dinamico a tempo continuo $dx/dt = F(x(t), u(t))$ produce traiettorie $\Phi(t, t_0, x_0)$ nello spazio degli stati. Il flusso $\{\Phi\}_{\geq 0}$ forma un semigruppo: $\Phi_{+2} =$

$\Phi \circ \Phi_2$. La distinzione tra sistema-funzione ($x(t) = f(t)$ determinato direttamente dal tempo) e sistema-flusso (la cui posizione è l'integrato della storia) è la distinzione formale tra assenza e presenza di spessore temporale. Il filtro di Kalman stima lo stato interno s da osservazioni rumorose attraverso un ciclo predizione-aggiornamento: $\hat{s}_{t|t-1} = A \cdot \hat{s}_{t-1|t-1} + B \cdot a$ (predizione); $\hat{s}_{t|t} = \hat{s}_{t|t-1} + K \cdot (s_t - C \cdot \hat{s}_{t|t-1})$ (aggiornamento), dove K è il guadagno di Kalman. Questo formalismo descrive la struttura del loop propriocettivo che mantiene una stima aggiornata dello stato corporeo.

AT.I.10 — Architetture JEPA e State Space Models

Una Joint Embedding Predictive Architecture (JEPA) è definita da: un encoder $s_2 = f\theta(x)$ che mappa osservazioni nello spazio latente; un predittore $\hat{s}_2 = g\varphi(s_2, z)$ che predice la rappresentazione latente di una seconda osservazione dato un vettore di contesto z; la perdita $\mathcal{L}_{2222}(\theta, \varphi) = \mathbb{E}[d(g\varphi(f\theta(x), z), f\theta(y))]$. La chiave filosofica è che la predizione avviene nello spazio latente astratto, non in quello delle osservazioni grezze. I modelli a spazio degli stati (SSM), come Mamba, hanno la forma: $h(t) = A \cdot h(t-1) + B \cdot x(t)$; $y(t) = C \cdot h(t)$, dove $h(t) \in \mathbb{R}^{\mathcal{D}}$ è uno stato nascosto che si aggiorna ricorrentemente. Gli autovalori di A con parte reale $\lambda_2 = -1/\tau_2$ determinano le costanti di tempo ritenzionali τ_2 del sistema, approssimando la struttura del flusso temporale genuino.

Appendice Tecnica II: Glossario dei Termini Tecnici

▪ Accoppiamento strutturale. In biologia dei sistemi (Maturana-Varela): la relazione co-adattativa tra un sistema autopoietico e il suo ambiente, in cui le perturbazioni ambientali selezionano tra i cambiamenti strutturali che il sistema può subire rimanendo vitale, senza determinarne direttamente la risposta.

▪ Architettura JEPA. Joint Embedding Predictive Architecture: sistema di apprendimento che predice rappresentazioni latenti di osservazioni future, anziché ricostruire le osservazioni nello spazio grezzo. Formulata da LeCun come candidata per l'intelligenza artificiale autonoma.

▪ Autopoiesi. Proprietà di sistemi biologici (Maturana-Varela) che generano e rigenerano i propri componenti attraverso la propria rete di processi, mantenendo una distinzione strutturale tra sé e ambiente attraverso un confine prodotto attivamente.

▪ Centroide di Fisher-Rao. Punto p^* di una varietà statistica che minimizza la somma delle distanze geodetiche al quadrato da un insieme di distribuzioni.

Utilizzato come indicatore del grado topologico di centrazione C(S) nella griglia dell'interno.

▪ Chiusura operazionale. Proprietà di un sistema autopoietico i cui processi interni formano una rete chiusa: ogni componente è prodotto dalla rete, e la rete è ciò che i componenti, nella loro interazione, producono e mantengono.

▪ Chiusura riflessiva. Proprietà di un sistema cognitivo che possiede stati aventi come contenuto i propri stati intenzionali. Formalmente: l'esistenza di una mappa M: Int(X) → States(X) che associa a ogni stato intenzionale uno stato avente quel primo stato come contenuto. Condizione necessaria per la coscienza nel senso delle teorie di ordine superiore.

▪ Coscienza fenomenica. La proprietà degli stati mentali di avere un carattere qualitativo in prima persona — un "com'è" (what-it-is-like-ness). Distinta dalla coscienza di accesso (Block) e dal punto di vista funzionale-strutturale dell'interno. Il suo rapporto con le condizioni strutturali è il nucleo del problema difficile della coscienza.

▪ Curvatura di Riemann. Tensore $R_{22}{}^{I}{}_{\square}$ che misura il fallimento della commutatività del trasporto parallelo su una varietà riemanniana. In geometria differenziale dell'incarnazione: curvatura non nulla nello spazio percettivo egocentrico riflette la distribuzione non uniforme della risoluzione sensoriale.

▪ Deficit di prospettiva. Vettore $\Delta(S) = (\Delta_1(S), ..., \Delta_5(S)) \in \mathbb{R}^{+}\mu$ che misura la distanza di un sistema S dal soddisfacimento di ciascuna delle cinque condizioni della griglia dell'interno. Indicatore sintetico della distanza strutturale dalla soglia dell'interno.

▪ Derivata covariante. Operazione di differenziazione ∇ su campi vettoriali di una varietà riemanniana, che generalizza la derivata ordinaria tenendo conto della curvatura dello spazio. I simboli di Christoffel $\Gamma_2{}^{I}{}_{\square}$ determinano come il trasporto parallelo "ruota" i vettori.

▪ Einfühlung. Termine husserliano per il processo pre-predicativo di accoppiamento analogico attraverso cui si attribuisce soggettività a un altro corpo percepito: base fenomenologica dell'intersoggettività e della teoria della mente.

▪ Energia libera variazionale. Quantità $F(o, q) = \mathbb{E}_{\mathbf{2}}[\ln q(\mu) - \ln p(o, \mu)]$ che funge da limite superiore alla log-evidenza negativa. La sua minimizzazione rispetto agli stati interni (percezione) e alle azioni (inferenza attiva) unifica percezione, azione e apprendimento in un unico principio (Friston).

▪ Equivalenza elementare. Relazione $\mathfrak{A} \equiv \mathfrak{B}$ tra strutture che soddisfano le stesse sentenze del primo ordine. Nucleo formale della prospettiva-neutralità: strutture elementarmente equivalenti sono indistinguibili da qualsiasi formula — nessun linguaggio formale può discernerle.

▪ Filtro generico (Cohen). Oggetto $G \subseteq P$ su un ordine parziale $\mathbf{L}$ che interseca ogni sottoinsieme denso del modello base M. Massimalmente anonimo: non è definibile da nessuna proprietà di M. Paradigma matematico dell'assenza di interno.

▪ Forcing. Tecnica sviluppata da Paul Cohen (1963) per costruire estensioni generiche di modelli della teoria degli insiemi e dimostrare risultati di indipendenza. La relazione di forcing $p \Vdash \varphi$ è definita interamente nel modello base M; il lemma di verità connette questa relazione alla verità nel modello generico M[G].

▪ Funzione propria. Nel senso di Ruth Millikan: la funzione per la quale un dispositivo biologico o cognitivo è stato selezionato nel corso dell'evoluzione o dell'apprendimento. Fonda il contenuto intenzionale: uno stato R rappresenta C perché la co-varianza tra R e C spiega, nella storia selettiva del sistema, perché quel tipo di stato è stato conservato.

▪ Grado di interno. Funzione $\mathbb{G}(S) = \prod_2 \mathbb{G}_2(S)^{\mathbf{X_2}} \in [0,1]$ aggregante i cinque indicatori parziali $\iota_2(S)$ della griglia dell'interno, con pesi w_2 che riflettono la gerarchia delle condizioni.

▪ Identità egocentrica vs. allocentrica. Distinzione tra sistemi di riferimento centrati sul corpo del soggetto (egocentrico) e sistemi di riferimento indipendenti dal corpo (allocentrico). La singolarità topologica dell'interno è l'origine del sistema egocentrico che non può essere arbitrariamente sostituita da un'altra.

▪ Informazione causale di Granger. $GC(X \to Y) = H(Y_{+1} \mid Y_1\square) - H(Y_{+1} \mid Y_1\square, X_1\square)$: riduzione di incertezza su Y prodotta dalla conoscenza del passato di X. Nell'indicatore IO(S, C): la versione condizionata che misura la causalità diretta di C su s al netto dell'intervento di agenti mediatori.

▪ Intenzionalità derivata/originale. Distinzione di Searle. Intenzionalità derivata: il contenuto di uno stato è determinato dall'interpretazione di agenti esterni dotati di intenzionalità originale. Intenzionalità originale: il contenuto è determinato dalla struttura interna e dalla storia selettiva propria del sistema, indipendentemente da interpretazioni esterne.

▪ Leib / Körper. Distinzione husserliana tra il corpo vissuto dall'interno come polo soggettivo dell'esperienza (Leib) e il corpo fisico oggettivo percepito dall'esterno (Körper). Solo il Leib introduce un Nullpunkt der Orientierung — l'origine assoluta di ogni sistema di riferimento egocentrico.

▪ Lemma di Yoneda. Risultato fondamentale della teoria delle categorie: per ogni funtore F: C → Set e ogni oggetto A, esiste una biiezione naturale Nat(hom(A,−), F) ≅ F(A). Implicazione: ogni oggetto categoriale è completamente determinato dal profilo dei suoi morfismi uscenti — identità come pura relazionalità esteriore.

▪ Manifold percettivo egocentrico. La varietà riemanniana (P, g***AEE***) con metrica non-omogenea che modellizza lo spazio percettivo di un sistema incarnato. La curvatura scalare è massima nelle regioni periferiche e minima nell'intorno dell'origine O: cattura la non-uniformità della risoluzione sensoriale.

▪ Metrica di Fisher. $g^{I}\square(\theta) = \mathbb{E}_{2}[\partial^{I} \log p(x;\theta) \cdot \partial \log p(x;\theta)]$: tensore riemanniano naturale sullo spazio dei parametri di una famiglia di distribuzioni di probabilità. Fonda la geometria dell'informazione e il grado topologico di centrazione C(S).

▪ Modello generativo. Distribuzione congiunta p(o, μ; θ) su osservazioni o e stati nascosti μ. In un sistema incarnato: p(o, $\mu_{22}\mathbf{r_{22}}$, $\mu_{22}\mathbb{I}\mathbf{r_{2}}\mathbf{rr}$) include $\mu_{22}\mathbb{I}\mathbf{r_{2}}\mathbf{rr}$ come variabile del modello: la prospettiva del sé è esplicita nella struttura generativa.

▪ Nullpunkt der Orientierung. "Punto zero dell'orientamento" (Husserl): l'origine assoluta del sistema di riferimento egocentrico corporeo, rispetto alla quale ogni altro punto è definito relazionalmente. Non è uno dei tanti sistemi di riferimento possibili: è la condizione di possibilità ontologica di qualsiasi altro sistema di riferimento.

▪ Punteggio deontico. In Brandom: il registro degli impegni e dei diritti inferenziali che i partecipanti a una pratica discorsiva mantengono reciprocamente. $\text{Score}_{2\to a}$: $\mathbb{P} \to \{-1, 0, +1\}$. La normatività culturale è costituita dalla partecipazione a questa struttura intersoggettiva di attribuzione reciproca di ragioni.

▪ Retenzionalità. Struttura della coscienza temporale (Husserl): la "coda" del presente vissuto che trattiene le impressioni immediatamente passate senza ridurle a ricordi espliciti. Formalizzata come operatore $R[x](\Delta) = k_{222}(\Delta) \cdot x(t+\Delta)$ con nucleo esponenziale $k_{222}(\Delta) = \exp(\Delta/\tau_{222})$.

▪ Secondo natura. In McDowell: la capacità acquisita attraverso la Bildung culturale di navigare lo spazio delle ragioni — di rispondere a fatti come a ragioni per credere o agire. Presupposto della normatività nel senso proposizionale pieno: non riducibile alla storia selettiva biologica, ma radicata in pratiche normative socialmente costituite.

▪ Simmetria spontaneamente rotta. Condizione di un sistema le cui leggi hanno simmetria G ma il cui stato fondamentale φ_0 ha solo simmetria $H \subset G$. Il punto φ_0

è il qui del sistema nello spazio dei minimi del potenziale: prima traccia formale dell'interno come rottura dell'omogeneità.

▪ Spazio di lavoro globale (GWT). Teoria della coscienza (Baars, Dehaene): l'accesso a un'informazione da parte dello spazio di lavoro globale — la sua diffusione a tutti i processori del sistema — costituisce la coscienza di accesso. Complementare a IIT: enfatizza la disponibilità globale anziché l'integrazione causale.

▪ Teoria IIT (Tononi). Integrated Information Theory: la coscienza è identica all'integrazione causale irriducibile di un sistema, misurata dalla quantità Φ. $\Phi(S, x) = \min_{\mathscr{P}} d(S, \mathscr{P})$, dove d è la distanza EMD tra strutture causa-effetto integrate e spezzate. $\Phi = 0$ implica nessuna coscienza; $\Phi > 0$ implica coscienza proporzionale.

▪ Trasparenza del PSM. Nel senso di Metzinger: l'auto-modello fenomenico è trasparente quando il sistema non può riconoscerlo come modello — quando il suo contenuto è vissuto come la realtà stessa piuttosto che come una rappresentazione di essa. La trasparenza è condizione della soggettività fenomenica e della sua fallibilità intrinseca.

Bibliografia

Filosofia della mente e coscienza

Baron-Cohen, S., Leslie, A. M. e Frith, U. (1985). Does the autistic child have a "theory of mind"? Cognition, 21(1), 37–46.

Block, N. (1995). On a Confusion about a Function of Consciousness. Behavioral and Brain Sciences, 18(2), 227–247.

Brandom, R. (1994). Making It Explicit: Reasoning, Representing, and Discursive Commitment. Harvard University Press.

Brandom, R. (2000). Articulating Reasons: An Introduction to Inferentialism. Harvard University Press.

Brentano, F. (1874). Psychologie vom empirischen Standpunkt. Duncker & Humblot.

Chalmers, D. J. (1995). Facing Up to the Problem of Conscious Experience. Journal of Consciousness Studies, 2(3), 200–219.

Chalmers, D. J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Chalmers, D. J. (2010). The Character of Consciousness. Oxford University Press.

Dennett, D. C. (1991). Consciousness Explained. Little, Brown and Company.

Dretske, F. (1981). Knowledge and the Flow of Information. MIT Press.

Dretske, F. (1988). Explaining Behavior: Reasons in a World of Causes. MIT Press.

Fodor, J. A. (1975). The Language of Thought. Harvard University Press.
Fodor, J. A. e Pylyshyn, Z. W. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. Cognition, 28(1–2), 3–71.
Frege, G. (1879). Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens. Louis Nebert.
Frege, G. (1892). Über Sinn und Bedeutung. Zeitschrift für Philosophie und philosophische Kritik, 100, 25–50.
Gallagher, S. (2005). How the Body Shapes the Mind. Oxford University Press.
Harnad, S. (1990). The Symbol Grounding Problem. Physica D: Nonlinear Phenomena, 42(1–3), 335–346.
Jackson, F. (1982). Epiphenomenal Qualia. The Philosophical Quarterly, 32(127), 127–136.
Kant, I. (1785). Grundlegung zur Metaphysik der Sitten. Johann Friedrich Hartknoch.
Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry e H. Wettstein (a cura di), Themes from Kaplan (pp. 481–563). Oxford University Press.
Korsgaard, C. M. (2009). Self-Constitution: Agency, Identity, and Integrity. Oxford University Press.
Kriegel, U. (2009). Subjective Consciousness: A Self-Representational Theory. Oxford University Press.
Loewer, B. e Rey, G. (a cura di) (1991). Meaning in Mind: Fodor and His Critics. Blackwell.
Lycan, W. G. (1996). Consciousness and Experience. MIT Press.
McDowell, J. (1994). Mind and World. Harvard University Press.
Metzinger, T. (2003). Being No One: The Self-Model Theory of Subjectivity. MIT Press.
Millikan, R. G. (1984). Language, Thought, and Other Biological Categories. MIT Press.
Millikan, R. G. (1993). White Queen Psychology and Other Essays for Alice. MIT Press.
Nagel, T. (1974). What Is It Like to Be a Bat? The Philosophical Review, 83(4), 435–450.
Ricoeur, P. (1983–1985). Temps et récit (3 voll.). Seuil.
Ricoeur, P. (1990). Soi-même comme un autre. Seuil.
Rosenthal, D. M. (1986). Two Concepts of Consciousness. Philosophical Studies, 49(3), 329–359.
Schechtman, M. (1996). The Constitution of Selves. Cornell University Press.
Searle, J. R. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences, 3(3), 417–424.
Searle, J. R. (1983). Intentionality: An Essay in the Philosophy of Mind. Cambridge University Press.
Wittgenstein, L. (1953). Philosophische Untersuchungen / Philosophical Investigations. Blackwell.
Zahavi, D. (2005). Subjectivity and Selfhood: Investigating the First-Person Perspective. MIT Press.

Fenomenologia

Husserl, E. (1900–1901). Logische Untersuchungen. Max Niemeyer.

Husserl, E. (1913). Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Max Niemeyer.

Husserl, E. (1929/1960). Cartesianische Meditationen. Hrsg. von S. Strasser. Martinus Nijhoff. (Trad. it.: Meditazioni cartesiane. Bompiani.)

Husserl, E. (1966). Zur Phänomenologie des inneren Zeitbewusstseins (1893–1917). Martinus Nijhoff. (Lezioni sulla coscienza interna del tempo).

Husserl, E. (1952). Ideen zu einer reinen Phänomenologie, Bd. II. Martinus Nijhoff.

James, W. (1890). The Principles of Psychology. Henry Holt.

Merleau-Ponty, M. (1945). Phénoménologie de la perception. Gallimard.

Matematica e logica

Amari, S. (1985). Differential-Geometrical Methods in Statistics. Springer.

Amari, S. e Nagaoka, H. (2000). Methods of Information Geometry. American Mathematical Society.

Bell, J. L. e Slomson, A. B. (1969). Models and Ultraproducts: An Introduction. North-Holland.

Cohen, P. J. (1963). The Independence of the Continuum Hypothesis. Proceedings of the National Academy of Sciences, 50(6), 1143–1148.

Cohen, P. J. (1966). Set Theory and the Continuum Hypothesis. W. A. Benjamin.

Kunen, K. (1980). Set Theory: An Introduction to Independence Proofs. North-Holland.

Mac Lane, S. (1971). Categories for the Working Mathematician. Springer.

Marker, D. (2002). Model Theory: An Introduction. Springer.

Tarski, A. (1936). Der Wahrheitsbegriff in den formalisierten Sprachen. Studia Philosophica, 1, 261–405.

Weyl, H. (1918). Das Kontinuum: Kritische Untersuchungen über die Grundlagen der Analysis. Veit.

Neuroscienze cognitive e scienze cognitive

Baker, C. L., Saxe, R. e Tenenbaum, J. B. (2009). Action Understanding as Inverse Planning. Cognition, 113(3), 329–349.

Baars, B. J. (1988). A Cognitive Theory of Consciousness. Cambridge University Press.

Dehaene, S. (2014). Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. Viking.

Dehaene, S., Changeux, J. P. e Naccache, L. (2011). The Global Neuronal Workspace Model of Conscious Access. In S. Dehaene e Y. Christen (a cura di), Characterizing Consciousness: From Cognition to the Clinic? Springer.

Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? Nature Reviews Neuroscience, 11(2), 127–138.

Friston, K. J. (2012). A Free Energy Principle for Biological Self-Organisation. PLOS Computational Biology, 8(11), e1002128.

Friston, K. J., Wiese, W. e Hobson, J. A. (2021). Sentience and the Minimal Self. In M. Tschantz, C. L. Buckley e A. K. Seth (a cura di), Active Inference. MIT Press.

Gallese, V., Fadiga, L., Fogassi, L. e Rizzolatti, G. (1996). Action Recognition in the Premotor Cortex. Brain, 119(2), 593–609.

Gallese, V. e Goldman, A. (1998). Mirror Neurons and the Simulation Theory of Mind-Reading. Trends in Cognitive Sciences, 2(12), 493–501.

O'Regan, J. K. e Noë, A. (2001). A Sensorimotor Account of Vision and Visual Consciousness. Behavioral and Brain Sciences, 24(5), 939–973.

Rao, R. P. N. e Ballard, D. H. (1999). Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects. Nature Neuroscience, 2(1), 79–87.

Rizzolatti, G. e Craighero, L. (2004). The Mirror-Neuron System. Annual Review of Neuroscience, 27, 169–192.

Tononi, G. (2004). An Information Integration Theory of Consciousness. BMC Neuroscience, 5, 42.

Tononi, G. (2014). Consciousness as Integrated Information: A Provisional Manifesto. Biological Bulletin, 215(3), 216–242.

Tomasello, M. (1995). Joint Attention as Social Cognition. In C. Moore e P. J. Dunham (a cura di), Joint Attention: Its Origins and Role in Development (pp. 103–130). Lawrence Erlbaum.

Tomasello, M. (2008). Origins of Human Communication. MIT Press.

Varela, F. J., Thompson, E. e Rosch, E. (1991). The Embodied Mind: Cognitive Science and Human Experience. MIT Press.

Wolpert, D. M. e Kawato, M. (1998). Multiple Paired Forward and Inverse Models for Motor Control. Neural Networks, 11(7–8), 1317–1329.

Biologia dei sistemi e teoria dell'evoluzione

Berg, H. C. e Brown, D. A. (1972). Chemotaxis in Escherichia coli Analysed by Three-Dimensional Tracking. Nature, 239(5374), 500–504.

Maturana, H. R. e Varela, F. J. (1980). Autopoiesis and Cognition: The Realization of the Living. D. Reidel.

Maynard Smith, J. e Szathmáry, E. (1995). The Major Transitions in Evolution. W. H. Freeman.

Thompson, E. (2007). Mind in Life: Biology, Phenomenology, and the Sciences of Mind. Harvard University Press.

Intelligenza artificiale e apprendimento automatico

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y. e Ballas, N. (2023). Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. Proceedings of CVPR 2023.

Brohan, A., Brown, N., Carbajal, J., et al. (2023). RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818.

Driess, D., Xia, F., Sajjadi, M. S. M., et al. (2023). PaLM-E: An Embodied Multimodal Language Model. arXiv:2303.03378.

Gu, A. e Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.

Hafner, D., Lillicrap, T., Ba, J. e Norouzi, M. (2019). Dream to Control: Learning Behaviors by Latent Imagination. arXiv:1912.01603.

Hafner, D., Lillicrap, T. P., Norouzi, M. e Ba, J. (2021). Mastering Atari with Discrete World Models. ICLR 2021.

Hafner, D., Pasukonis, J., Ba, J. e Lillicrap, T. (2023). Mastering Diverse Domains through World Models. arXiv:2301.04104.

Hansen, N., Wang, X. e Su, H. (2023). TD-MPC2: Scalable, Robust World Models for Continuous Control. arXiv:2310.16828.

Hochreiter, S. e Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.

Huang, Y. e Raza, H. (2026). Knowledge, Rules and Their Embeddings: Two Paths towards Neuro-Symbolic JEPA. Preprint HAL hal-05527907.

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. OpenReview.net preprint.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. e Klimov, O. (2017). Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B. e Liu, Y. (2021). RoFormer: Enhanced Transformer with Rotary Position Embedding. arXiv:2104.09864.

Sutton, R. S. e Barto, A. G. (1998). Reinforcement Learning: An Introduction. MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. e Polosukhin, I. (2017). Attention Is All You Need. NeurIPS 2017.

Fisica teorica e teoria dei sistemi complessi

Goldenfeld, N. (1992). Lectures on Phase Transitions and the Renormalization Group. Addison-Wesley.

Kauffman, S. A. (1993). The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press.

Klein, F. (1872). Vergleichende Betrachtungen über neuere geometrische Forschungen (Erlanger Programm). Deichert.

Penrose, R. (1989). The Emperor's New Mind. Oxford University Press.